EE219: Large-Scale Data Mining: Models and Algorithms
Winter 2018

Project 5: Popularity Prediction on Twitter

Akshya ARUNACHALAM (UID: 904-943-191)
Zhi Ming CHUA (UID: 805-068-401)
Ashwin Kumar KANNAN (UID: 605-035-204)
Vijay RAVI (UID: 805-033-666)

March 19, 2018

Contents

1 Introduction

Twitter is one of the most commonly used social networking services and it serves as a platform for public discussion. Hashtags are used to tag tweets, allowing users to easily find posts with a specific theme or content. Using this, analysis can be performed on tweet activities of different hashtags, including predicting the future popularity of a particular topic.

For this project, the Twitter data used is a collection of tweets with certain popular hashtags from 2 weeks prior to and 1 week after Super Bowl 2015, and linear regression models are used to make predictions for other hashtags based on this dataset.

2 Popularity Prediction

2.1 Training Data

The following statistics for each hashtag is calculated for the training and tabulated in Table 1.

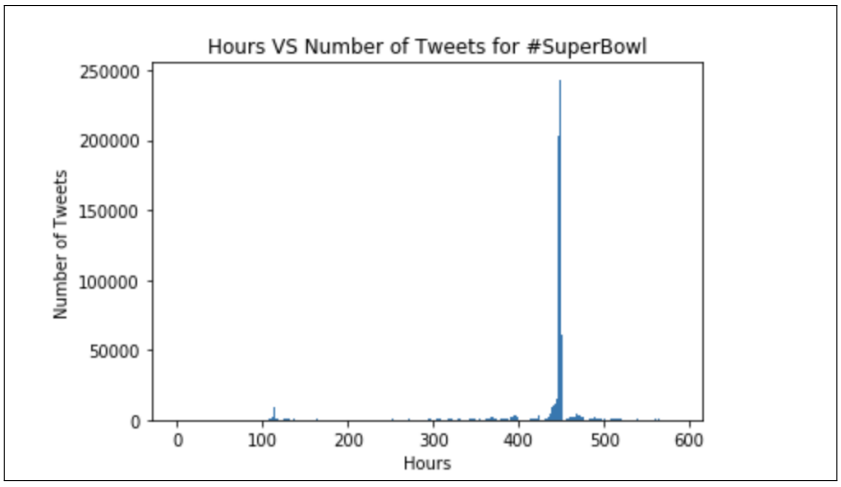
- Average number of tweets per hour
- Average number of followers of users posting the tweets
- Average number of retweets

Hashtag	**#NFL**	**#SuperBowl**	**#SB49**	**#GoHawks**	**#GoPatriots**	**#Patriots**
Average number of tweets per hour	441.3	2303	142.0	325.4	45.69	83.46
Average number of followers of user posting the tweets	4290	3592	2235	1545	1299	1650
Average number of retweets	1.539	2.388	2.511	2.015	1.400	1.783

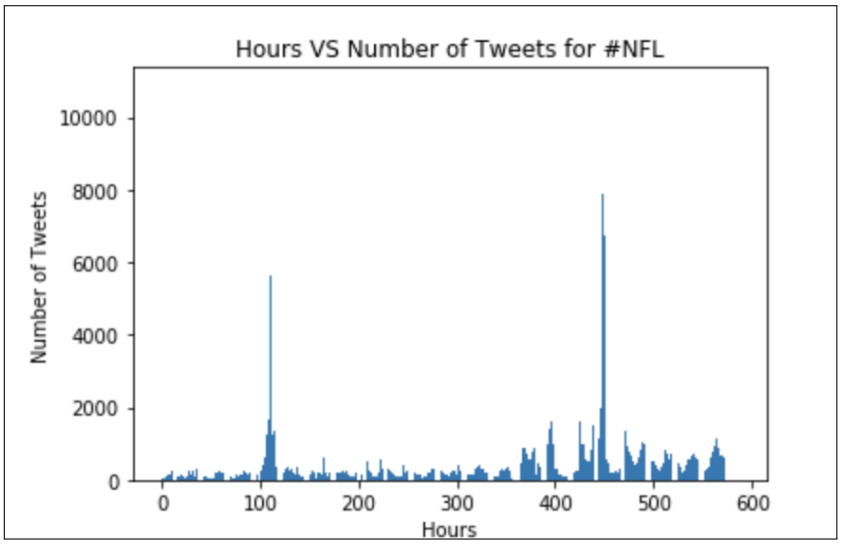
Table 1: Training Data Statistics

#SuperBowl, #NFL and #SB49 had the highest average number of tweets per hour, average number of average followers for users that posted the tweets and average number of retweets, respectively. The higher statistics of these hashtags could be attributed to the fact that they are used by the users tweeting about the event in general, whereas **#GoHawks** and **#GoPatriots** are team-specific hashtags, which are mostly used only by the respective teams' fans.

The number of tweets per hour for #NFL and #SuperBowl are plotted as a histogram with 1-hour bins in Figure 1.



(a) #SuperBowl



(b) #NFL

Figure 1: Number of Tweets Against Hour

Two peaks can be observed in the histogram plots for both hashtags. The second peak is higher, suggesting that it might be when the SuperBowl took place. The first peak is approximately two weeks before the SuperBowl and could be when the finalist teams were confirmed.

2.2 Linear Regression Model

In this section, a separate linear regression model is used to predict the number of tweets in the next hour for each of the five hashtags. The features used are:

1. Number of tweets
2. Total number of retweets
3. Sum of the number of followers of the users posting the hashtag
4. Maximum number of followers of the users posting the hashtag
5. Time of the day (one-hot encoded)

The following statistics are used to analyze the results from each model.

- p-Value:
  The p-value is a number from 0 to 1, and it weighs the strength of the evidence against the null hypothesis. If the p-value is lower than the level of significance, it can be interpreted as the data providing sufficiently strong evidence to reject the null hypotheses and that there is high correlation between the dependent and independent variables.
- t-Test :
  The t-test can be used to find evidence of a significant difference between the population mean and a hypothesized value. The t-value is calculated in units of standard error and the greater the magnitude, the greater the evidence against the null hypothesis that there is no significant difference.
- R-Squared :
  R-squared is a statistical measure of how closely the data is fitted to the regression line and its value ranges between 0% and 100%. The higher the value, the better the model does in explaining the variability around the mean.
- Root-Mean-Square Error (RMSE):
  The RMSE is used to quantify the deviation of the predicted values from the actual values and hence is a measure of accuracy. As the effect of each error is proportional to the squared error, larger errors have a disproportionately large effect on the RMSE.

For each hastag, the predicted number of tweets in the next hour is plotted against the actual number of tweets. The ordinary least squares (OLS) linear regression model results are also reported.

2.2.1 #NFL

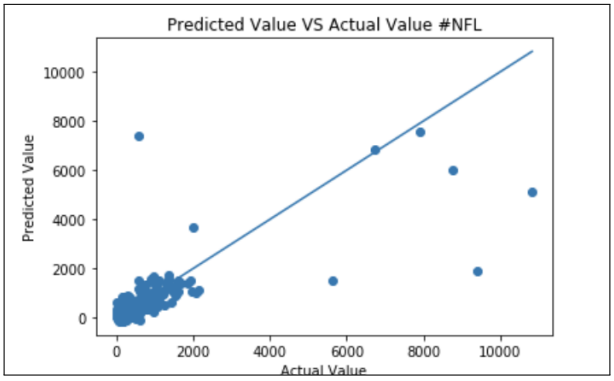


Figure 2: Predicted Number of Tweets Against Actual Number of Tweets

OLS Regression Results

Dep. Variable:	target_value	R-squared:	0.585
Model:	OLS	Adj. R-squared:	0.565
Method:	Least Squares	F-statistic:	29.17
Date:	Thu, 08 Mar 2018	Prob (F-statistic):	7.38e-89
Time:	18:46:30	Log-Likelihood:	-4554.8
No. Observations:	587	AIC:	9166.
Df Residuals:	559	BIC:	9288.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
hourly_tweet_number	0.4752	0.103	4.601	0.000	0.272	0.678
hourly_retweet_number	-1.5832	2.807	-0.564	0.573	-7.096	3.930
hourly_sum_of_followers	7.315e-05	2.65e-05	2.756	0.006	2.1e-05	0.000
hourly_max_follower	-9.432e-05	3.7e-05	-2.547	0.011	-0.000	-2.16e-05
time_of_day_0.0	56.3083	116.806	0.482	0.630	-173.124	285.741
time_of_day_1.0	67.4735	116.873	0.577	0.564	-162.091	297.038
time_of_day_2.0	76.3124	116.595	0.655	0.513	-152.705	305.330
time_of_day_3.0	92.3850	116.643	0.792	0.429	-136.727	321.497
time_of_day_4.0	83.8884	116.783	0.718	0.473	-145.500	313.276
time_of_day_5.0	169.5487	116.801	1.452	0.147	-59.874	398.971
time_of_day_6.0	196.7151	117.237	1.678	0.094	-33.563	426.993
time_of_day_7.0	149.7344	117.061	1.279	0.201	-80.199	379.668
time_of_day_8.0	166.4218	117.790	1.413	0.158	-64.943	397.787
time_of_day_9.0	120.6821	117.350	1.028	0.304	-109.818	351.182
time_of_day_10.0	237.9457	117.583	2.024	0.043	6.986	468.905
time_of_day_11.0	244.0414	120.708	2.022	0.044	6.945	481.138
time_of_day_12.0	83.3442	120.386	0.692	0.489	-153.121	319.809
time_of_day_13.0	220.2633	121.927	1.807	0.071	-19.229	459.755
time_of_day_14.0	634.0510	121.109	5.235	0.000	396.167	871.934
time_of_day_15.0	15.7214	123.964	0.127	0.899	-227.771	259.213
time_of_day_16.0	269.0316	122.193	2.202	0.028	29.018	509.045
time_of_day_17.0	218.6963	125.524	1.742	0.082	-27.860	465.253
time_of_day_18.0	283.3371	122.493	2.313	0.021	42.734	523.940
time_of_day_19.0	-198.0234	122.052	-1.622	0.105	-437.760	41.713
time_of_day_20.0	86.8392	120.082	0.723	0.470	-149.027	322.706
time_of_day_21.0	81.4536	119.888	0.679	0.497	-154.032	316.939
time_of_day_22.0	33.0973	119.432	0.277	0.782	-201.494	267.688
time_of_day_23.0	63.9559	119.234	0.536	0.592	-170.245	298.156

Omnibus:	581.889	Durbin-Watson:	2.364
Prob(Omnibus):	0.000	Jarque-Bera (JB):	286567.909
Skew:	3.503	Prob(JB):	0.00
Kurtosis:	111.016	Cond. No.	2.97e+07

Figure 3: OLS Regression Results for #NFL

The RMSE is 567.1 and the R-squared value is 0.585.

2.2.2 #SuperBowl

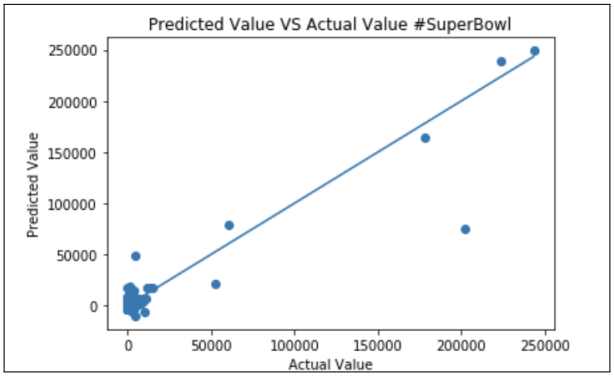


Figure 4: Predicted Number of Tweets Against Actual Number of Tweets

OLS Regression Results

Dep. Variable:	target_value	R-squared:	0.870
Model:	OLS	Adj. R-squared:	0.864
Method:	Least Squares	F-statistic:	138.6
Date:	Thu, 08 Mar 2018	Prob (F-statistic):	1.86e-227
Time:	18:52:04	Log-Likelihood:	-5970.3
No. Observations:	586	AIC:	1.200e+04
Df Residuals:	558	BIC:	1.212e+04
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
hourly_tweet_number	2.2978	0.084	27.221	0.000	2.132	2.464
hourly_retweet_number	-3.9616	1.741	-2.275	0.023	-7.381	-0.542
hourly_sum_of_followers	-0.0002	1.2e-05	-19.193	0.000	-0.000	-0.000
hourly_max_follower	0.0011	0.000	9.488	0.000	0.001	0.001
time_of_day_0.0	-1152.9726	1329.978	-0.867	0.386	-3765.347	1459.402
time_of_day_1.0	-523.1336	1320.899	-0.396	0.692	-3117.676	2071.408
time_of_day_2.0	-320.5041	1321.701	-0.242	0.808	-2916.622	2275.614
time_of_day_3.0	-754.8762	1327.757	-0.569	0.570	-3362.888	1853.136
time_of_day_4.0	-325.6119	1322.931	-0.246	0.806	-2924.145	2272.921
time_of_day_5.0	-656.1878	1335.849	-0.491	0.623	-3280.096	1967.720
time_of_day_6.0	-210.0972	1330.078	-0.158	0.875	-2822.669	2402.474
time_of_day_7.0	-707.7970	1337.028	-0.529	0.597	-3334.019	1918.425
time_of_day_8.0	-50.0254	1340.333	-0.037	0.970	-2682.741	2582.690
time_of_day_9.0	-659.1035	1337.665	-0.493	0.622	-3286.578	1968.371
time_of_day_10.0	-954.4856	1376.758	-0.693	0.488	-3658.747	1749.776
time_of_day_11.0	-460.5540	1373.900	-0.335	0.738	-3159.202	2238.094
time_of_day_12.0	-182.9047	1365.986	-0.134	0.894	-2866.008	2500.198
time_of_day_13.0	981.7322	1375.887	0.714	0.476	-1720.818	3684.283
time_of_day_14.0	5249.6969	1390.505	3.775	0.000	2518.433	7980.961
time_of_day_15.0	-2006.0965	1395.997	-1.437	0.151	-4748.148	735.955
time_of_day_16.0	-1450.4399	1388.073	-1.045	0.297	-4176.926	1276.047
time_of_day_17.0	24.4359	1389.325	0.018	0.986	-2704.509	2753.381
time_of_day_18.0	-2880.8218	1370.936	-2.101	0.036	-5573.647	-187.997
time_of_day_19.0	-3352.1905	1369.294	-2.448	0.015	-6041.791	-662.590
time_of_day_20.0	-1327.1132	1361.904	-0.974	0.330	-4002.199	1347.972
time_of_day_21.0	-1105.3829	1355.967	-0.815	0.415	-3768.806	1558.040
time_of_day_22.0	-799.4634	1351.176	-0.592	0.554	-3453.476	1854.549
time_of_day_23.0	-1423.8574	1358.102	-1.048	0.295	-4091.475	1243.760

Omnibus:	1121.631	Durbin-Watson:	1.721
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1763400.781
Skew:	12.869	Prob(JB):	0.00
Kurtosis:	270.505	Cond. No.	7.60e+08

Figure 5: OLS Regression Results for #SuperBowl

The RMSE is 6433 and the R-squared value is 0.870, suggesting that the model is well-fitted.

2.2.3 #SB49

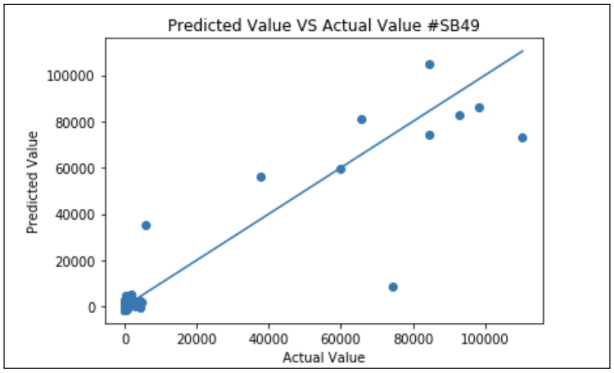


Figure 6: Predicted Number of Tweets Against Actual Number of Tweets

OLS Regression Results

Dep. Variable:	target_value	R-squared:	0.857
Model:	OLS	Adj. R-squared:	0.850
Method:	Least Squares	F-statistic:	122.7
Date:	Thu, 08 Mar 2018	Prob (F-statistic):	3.81e-214
Time:	18:49:04	Log-Likelihood:	-5632.7
No. Observations:	583	AIC:	1.132e+04
Df Residuals:	555	BIC:	1.144e+04
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
hourly_tweet_number	0.9414	0.030	31.904	0.000	0.883	0.999
hourly_retweet_number	-75.1845	10.554	-7.124	0.000	-95.916	-54.453
hourly_sum_of_followers	3.355e-06	4.49e-06	0.748	0.455	-5.46e-06	1.22e-05
hourly_max_follower	0.0002	4.15e-05	4.185	0.000	9.23e-05	0.000
time_of_day_0.0	-70.8379	780.105	-0.091	0.928	-1603.158	1461.482
time_of_day_1.0	-102.1442	782.115	-0.131	0.896	-1638.412	1434.124
time_of_day_2.0	-164.2185	782.418	-0.210	0.834	-1701.082	1372.645
time_of_day_3.0	-197.2752	782.395	-0.252	0.801	-1734.093	1339.542
time_of_day_4.0	-256.9267	786.731	-0.327	0.744	-1802.261	1288.407
time_of_day_5.0	2494.5636	785.829	3.174	0.002	951.000	4038.127
time_of_day_6.0	1319.3125	785.055	1.681	0.093	-222.729	2861.354
time_of_day_7.0	-1295.7993	807.914	-1.604	0.109	-2882.742	291.144
time_of_day_8.0	-1039.8350	803.523	-1.294	0.196	-2618.153	538.483
time_of_day_9.0	30.3122	819.470	0.037	0.971	-1579.329	1639.953
time_of_day_10.0	-182.7876	809.576	-0.226	0.821	-1772.996	1407.421
time_of_day_11.0	205.5069	804.837	0.255	0.799	-1375.392	1786.406
time_of_day_12.0	-430.8907	817.501	-0.527	0.598	-2036.664	1174.883
time_of_day_13.0	-1031.0163	804.730	-1.281	0.201	-2611.704	549.672
time_of_day_14.0	-1397.9380	797.865	-1.752	0.080	-2965.143	169.267
time_of_day_15.0	-65.5140	799.750	-0.082	0.935	-1636.421	1505.393
time_of_day_16.0	-50.2012	799.358	-0.063	0.950	-1620.339	1519.936
time_of_day_17.0	-292.5488	803.164	-0.364	0.716	-1870.161	1285.063
time_of_day_18.0	-59.0724	795.882	-0.074	0.941	-1622.381	1504.237
time_of_day_19.0	-174.8924	797.157	-0.219	0.826	-1740.706	1390.921
time_of_day_20.0	-136.8856	795.658	-0.172	0.863	-1699.754	1425.983
time_of_day_21.0	-6.9382	795.166	-0.009	0.993	-1568.841	1554.965
time_of_day_22.0	40.7550	795.142	0.051	0.959	-1521.100	1602.610
time_of_day_23.0	41.0264	795.259	0.052	0.959	-1521.059	1603.112

Omnibus:	945.198	Durbin-Watson:	1.327
Prob(Omnibus):	0.000	Jarque-Bera (JB):	749151.776
Skew:	9.166	Prob(JB):	0.00
Kurtosis:	177.653	Cond. No.	4.82e+08

Figure 7: OLS Regression Results for #SB49

The RMSE is 3800 and the R-squared value is 0.857, suggesting that it is a well-fitting model.

2.2.4 #GoHawks

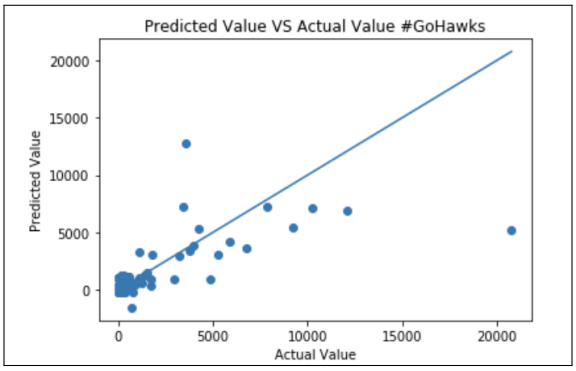


Figure 8: Predicted Number of Tweets Against Actual Number of Tweets

OLS Regression Results

Dep. Variable:	target_value	R-squared:	0.532
Model:	OLS	Adj. R-squared:	0.509
Method:	Least Squares	F-statistic:	23.19
Date:	Thu, 08 Mar 2018	Prob (F-statistic):	1.44e-73
Time:	18:43:46	Log-Likelihood:	-4767.2
No. Observations:	579	AIC:	9590.
Df Residuals:	551	BIC:	9712.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
hourly_tweet_number	1.4237	0.164	8.659	0.000	1.101	1.747
hourly_retweet_number	-19.8237	3.889	-5.097	0.000	-27.463	-12.184
hourly_sum_of_followers	-0.0003	8.51e-05	-3.270	0.001	-0.000	-0.000
hourly_max_follower	0.0002	0.000	1.126	0.261	-0.000	0.001
time_of_day_0.0	16.9938	187.031	0.091	0.928	-350.387	384.375
time_of_day_1.0	12.1762	186.821	0.065	0.948	-354.793	379.145
time_of_day_2.0	3.2168	186.790	0.017	0.986	-363.691	370.124
time_of_day_3.0	23.4880	190.666	0.123	0.902	-351.034	398.010
time_of_day_4.0	34.2407	190.662	0.180	0.858	-340.273	408.754
time_of_day_5.0	52.2219	190.884	0.274	0.785	-322.727	427.171
time_of_day_6.0	120.2952	190.978	0.630	0.529	-254.838	495.428
time_of_day_7.0	142.3044	190.957	0.745	0.456	-232.788	517.397
time_of_day_8.0	205.6231	193.187	1.064	0.288	-173.850	585.096
time_of_day_9.0	197.4692	191.701	1.030	0.303	-179.085	574.023
time_of_day_10.0	324.2885	191.838	1.690	0.092	-52.535	701.112
time_of_day_11.0	215.3796	191.699	1.124	0.262	-161.170	591.929
time_of_day_12.0	17.7893	192.423	0.092	0.926	-360.182	395.761
time_of_day_13.0	253.5218	192.556	1.317	0.189	-124.711	631.755
time_of_day_14.0	999.8690	193.175	5.176	0.000	620.420	1379.318
time_of_day_15.0	-243.6088	198.152	-1.229	0.219	-632.835	145.618
time_of_day_16.0	199.9072	193.022	1.036	0.301	-179.241	579.055
time_of_day_17.0	172.4385	193.656	0.890	0.374	-207.955	552.832
time_of_day_18.0	-130.1525	192.812	-0.675	0.500	-508.889	248.583
time_of_day_19.0	-16.4555	197.652	-0.083	0.934	-404.699	371.788
time_of_day_20.0	157.2451	193.179	0.814	0.416	-222.213	536.703
time_of_day_21.0	85.4406	193.508	0.442	0.659	-294.662	465.543
time_of_day_22.0	26.5137	191.449	0.138	0.890	-349.545	402.572
time_of_day_23.0	-7.8748	190.873	-0.041	0.967	-382.803	367.053

Omnibus:	854.260	Durbin-Watson:	2.193
Prob(Omnibus):	0.000	Jarque-Bera (JB):	683918.794
Skew:	7.479	Prob(JB):	0.00
Kurtosis:	170.706	Cond. No.	1.62e+07

Figure 9: OLS Regression Results for #GoHawks

The RMSE is 911.0 and the R-squared value is 0.532.

2.2.5 #GoPatriots

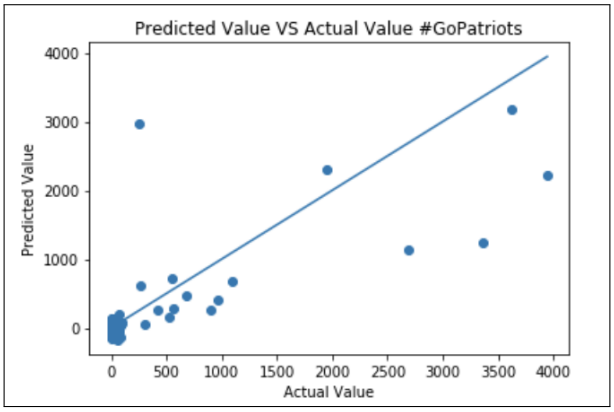


Figure 10: Predicted Number of Tweets Against Actual Number of Tweets

OLS Regression Results

Dep. Variable:	target_value	R-squared:	0.627
Model:	OLS	Adj. R-squared:	0.608
Method:	Least Squares	F-statistic:	34.00
Date:	Thu, 08 Mar 2018	Prob (F-statistic):	6.50e-99
Time:	18:44:56	Log-Likelihood:	-3827.5
No. Observations:	575	AIC:	7711.
Df Residuals:	547	BIC:	7833.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
hourly_tweet_number	0.2511	0.203	1.236	0.217	-0.148	0.650
hourly_retweet_number	-10.9993	3.836	-2.867	0.004	-18.535	-3.464
hourly_sum_of_followers	0.0006	0.000	3.230	0.001	0.000	0.001
hourly_max_follower	-0.0007	0.000	-3.548	0.000	-0.001	-0.000
time_of_day_0.0	3.1339	39.392	0.080	0.937	-74.245	80.512
time_of_day_1.0	3.7415	39.391	0.095	0.924	-73.634	81.117
time_of_day_2.0	6.1516	39.394	0.156	0.876	-71.231	83.535
time_of_day_3.0	9.5014	39.413	0.241	0.810	-67.918	86.921
time_of_day_4.0	10.9079	39.429	0.277	0.782	-66.543	88.359
time_of_day_5.0	11.4425	39.409	0.290	0.772	-65.970	88.855
time_of_day_6.0	9.7617	39.410	0.248	0.804	-67.651	87.174
time_of_day_7.0	20.9183	39.428	0.531	0.596	-56.531	98.368
time_of_day_8.0	15.4460	39.403	0.392	0.695	-61.953	92.845
time_of_day_9.0	23.8400	39.557	0.603	0.547	-53.862	101.542
time_of_day_10.0	14.1391	39.437	0.359	0.720	-63.328	91.606
time_of_day_11.0	33.6158	40.157	0.837	0.403	-45.265	112.497
time_of_day_12.0	119.8666	39.450	3.038	0.002	42.374	197.359
time_of_day_13.0	49.6551	39.945	1.243	0.214	-28.810	128.120
time_of_day_14.0	-22.2959	40.144	-0.555	0.579	-101.151	56.559
time_of_day_15.0	112.0844	39.919	2.808	0.005	33.671	190.497
time_of_day_16.0	80.0945	39.711	2.017	0.044	2.089	158.100
time_of_day_17.0	-122.2721	39.740	-3.077	0.002	-200.335	-44.210
time_of_day_18.0	13.7698	39.708	0.347	0.729	-64.229	91.769
time_of_day_19.0	7.9766	39.415	0.202	0.840	-69.447	85.400
time_of_day_20.0	7.8640	39.419	0.199	0.842	-69.567	85.295
time_of_day_21.0	6.7570	39.440	0.171	0.864	-70.715	84.229
time_of_day_22.0	2.7549	39.393	0.070	0.944	-74.625	80.135
time_of_day_23.0	2.9905	40.239	0.074	0.941	-76.051	82.032

Omnibus:	272.512	Durbin-Watson:	2.136
Prob(Omnibus):	0.000	Jarque-Bera (JB):	358447.958
Skew:	-0.297	Prob(JB):	0.00
Kurtosis:	125.315	Cond. No.	2.16e+06

Figure 11: OLS Regression Results for #GoPatriots

The RMSE is 188.2 and the R-Squared value is 0.627.

2.2.6 #Patriots

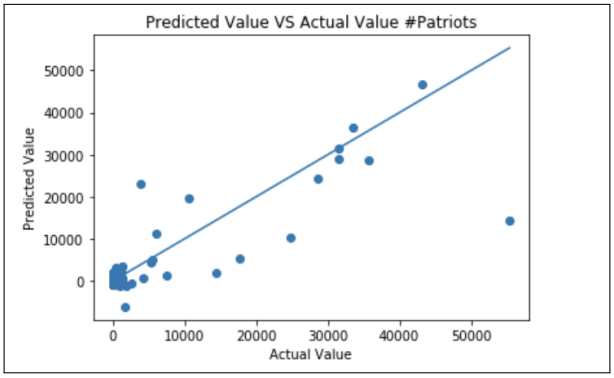


Figure 12: Predicted Number of Tweets Against Actual Number of Tweets

OLS Regression Results

Dep. Variable:	target_value	R-squared:	0.723
Model:	OLS	Adj. R-squared:	0.709
Method:	Least Squares	F-statistic:	53.98
Date:	Thu, 08 Mar 2018	Prob (F-statistic):	9.92e-137
Time:	18:47:30	Log-Likelihood:	-5377.2
No. Observations:	587	AIC:	1.081e+04
Df Residuals:	559	BIC:	1.093e+04
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
hourly_tweet_number	0.8230	0.039	21.051	0.000	0.746	0.900
hourly_retweet_number	-19.7359	3.548	-5.562	0.000	-26.706	-12.766
hourly_sum_of_followers	3.505e-05	2.54e-05	1.382	0.168	-1.48e-05	8.49e-05
hourly_max_follower	0.0001	9.69e-05	1.402	0.161	-5.45e-05	0.000
time_of_day_0.0	59.0273	472.067	0.125	0.901	-868.216	986.270
time_of_day_1.0	47.4428	471.927	0.101	0.920	-879.524	974.409
time_of_day_2.0	50.7466	471.983	0.108	0.914	-876.330	977.823
time_of_day_3.0	77.6890	472.291	0.164	0.869	-849.993	1005.371
time_of_day_4.0	79.4322	473.004	0.168	0.867	-849.650	1008.515
time_of_day_5.0	76.4594	474.558	0.161	0.872	-855.676	1008.594
time_of_day_6.0	70.7354	473.728	0.149	0.881	-859.769	1001.240
time_of_day_7.0	21.2432	476.106	0.045	0.964	-913.932	956.419
time_of_day_8.0	96.7734	473.981	0.204	0.838	-834.228	1027.774
time_of_day_9.0	500.0589	474.386	1.054	0.292	-431.738	1431.856
time_of_day_10.0	1805.4470	474.871	3.802	0.000	872.696	2738.198
time_of_day_11.0	-44.4154	487.337	-0.091	0.927	-1001.651	912.820
time_of_day_12.0	-306.7086	489.203	-0.627	0.531	-1267.609	654.192
time_of_day_13.0	31.0503	488.518	0.064	0.949	-928.505	990.606
time_of_day_14.0	435.8040	491.708	0.886	0.376	-530.017	1401.625
time_of_day_15.0	-78.9151	492.324	-0.160	0.873	-1045.946	888.115
time_of_day_16.0	-463.0496	498.382	-0.929	0.353	-1441.979	515.880
time_of_day_17.0	1203.5731	486.460	2.474	0.014	248.061	2159.085
time_of_day_18.0	-92.8105	490.764	-0.189	0.850	-1056.778	871.157
time_of_day_19.0	-841.7744	484.449	-1.738	0.083	-1793.337	109.788
time_of_day_20.0	271.5677	491.369	0.553	0.581	-693.587	1236.722
time_of_day_21.0	68.8244	483.349	0.142	0.887	-880.578	1018.226
time_of_day_22.0	95.4799	482.402	0.198	0.843	-852.062	1043.022
time_of_day_23.0	28.7491	482.046	0.060	0.952	-918.094	975.593

Omnibus:	974.392	Durbin-Watson:	1.943
Prob(Omnibus):	0.000	Jarque-Bera (JB):	835105.074
Skew:	9.619	Prob(JB):	0.00
Kurtosis:	186.777	Cond. No.	5.27e+07

Figure 13: OLS Regression Results for #Patriots

The RMSE is 2302 and the R-squared value is 0.723.

2.2.7 Analysis

The first four features are generally significant features for different hashtags. An interesting observation made is that the 14^{th} hour of the day is a significant feature, with much larger t-values and much smaller P-values compared to the other hours of the day.

2.3 Additional Features

In this section, additional features are used to predict the number of tweets in the next hour using a linear regression model. The 7 additional features are as follows:

1. **User mentions:** Number of mentions of a tweet
2. **Number of hashtags in tweet:** Number of hashtags in a tweet may affect its popularity and visibility
3. **Number of replies:** A tweet that receives a reply will also appear in the feed of the followers of the user that replied
4. **Number of favorites:** Number of times a tweet has been favorited may be an indication of how popular the opinion is
5. **Ranking scores**
6. **Number of citations:** Higher number of citations indicate a higher popularity of the tweet
7. **Verified User or not:** Verified users typically have more followers, making their tweets more favorable

These features are user dependent and may have a direct impact on the popularity of the hashtag, therefore helping to predict the number of tweets in the next hour.

Using these features in conjunction with those identified in the previous section, a total of 12 features are used to fit the linear regression model. For each hashtag, the fitting accuracy and significance of variables are reported. Scatter plots of predictants against the values of the top 3 most significant features are also shown in the subsequent sections.

2.3.1 #NFL

OLS Regression Results

Dep. Variable:	target_value	R-squared:	0.745
Model:	OLS	Adj. R-squared:	0.730
Method:	Least Squares	F-statistic:	50.47
Date:	Fri, 09 Mar 2018	Prob (F-statistic):	4.04e-142
Time:	16:49:18	Log-Likelihood:	-4412.3
No. Observations:	587	AIC:	8891.
Df Residuals:	554	BIC:	9035.
Df Model:	32		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
hourly_tweet_number	-0.2686	0.787	-0.341	0.733	-1.814	1.277
hourly_retweet_number	-6.0238	2.555	-2.357	0.019	-11.043	-1.005
hourly_sum_of_followers	-3.224e-05	2.54e-05	-1.268	0.205	-8.22e-05	1.77e-05
hourly_max_follower	4.066e-05	3.4e-05	1.194	0.233	-2.62e-05	0.000
hourly_user_mentions	1.6992	0.652	2.604	0.009	0.418	2.981
hourly_hashtags_in_tweet	0.6578	0.098	6.698	0.000	0.465	0.851
hourly_reply	-0.2686	0.787	-0.341	0.733	-1.814	1.277
hourly_favorite_count	-1.9212	0.182	-10.546	0.000	-2.279	-1.563
hourly_verified	-2.142e-07	2.66e-07	-0.807	0.420	-7.36e-07	3.07e-07
hourly_number_of_citations	0.4617	0.140	3.290	0.001	0.186	0.737
hourly_ranking_scores	-0.2054	0.321	-0.640	0.522	-0.836	0.425
time_of_day_0.0	-69.7005	95.015	-0.734	0.464	-256.335	116.934
time_of_day_1.0	-71.8541	94.259	-0.762	0.446	-257.002	113.294
time_of_day_2.0	-38.1149	93.430	-0.408	0.683	-221.635	145.406
time_of_day_3.0	-14.8486	93.182	-0.159	0.873	-197.882	168.185
time_of_day_4.0	8.7232	93.466	0.093	0.926	-174.868	192.314
time_of_day_5.0	78.9256	93.196	0.847	0.397	-104.135	261.986
time_of_day_6.0	81.7231	94.542	0.864	0.388	-103.982	267.428
time_of_day_7.0	3.2689	96.142	0.034	0.973	-185.578	192.116
time_of_day_8.0	-31.3222	98.531	-0.318	0.751	-224.863	162.219
time_of_day_9.0	39.6039	98.241	0.403	0.687	-153.366	232.574
time_of_day_10.0	57.8447	97.448	0.594	0.553	-133.567	249.257
time_of_day_11.0	10.4495	103.138	0.101	0.919	-192.140	213.039
time_of_day_12.0	-135.5008	100.882	-1.343	0.180	-333.658	62.657
time_of_day_13.0	-41.6154	101.186	-0.411	0.681	-240.370	157.139
time_of_day_14.0	368.6297	101.311	3.639	0.000	169.628	567.631
time_of_day_15.0	-256.1377	100.637	-2.545	0.011	-453.815	-58.461
time_of_day_16.0	19.1353	100.595	0.190	0.849	-178.459	216.730
time_of_day_17.0	28.6657	102.894	0.279	0.781	-173.444	230.775
time_of_day_18.0	50.2220	101.113	0.497	0.620	-148.391	248.835
time_of_day_19.0	-141.7708	99.502	-1.425	0.155	-337.218	53.676
time_of_day_20.0	-112.5821	98.655	-1.141	0.254	-306.366	81.202
time_of_day_21.0	-76.4407	99.061	-0.772	0.441	-271.021	118.140
time_of_day_22.0	-145.6697	97.391	-1.496	0.135	-336.970	45.631
time_of_day_23.0	-87.7202	97.371	-0.901	0.368	-278.982	103.541

Omnibus:	684.853	Durbin-Watson:	2.578
Prob(Omnibus):	0.000	Jarque-Bera (JB):	94183.605
Skew:	5.330	Prob(JB):	0.00
Kurtosis:	64.132	Cond. No.	1.82e+20

Figure 14: OLS Regression Results for #NFL

The RMSE is 444.9 and the R-squared value is 0.745.

2.3.2 #SuperBowl

OLS Regression Results

Dep. Variable:	target_value	R-squared:	0.909
Model:	OLS	Adj. R-squared:	0.904
Method:	Least Squares	F-statistic:	173.0
Date:	Fri, 09 Mar 2018	Prob (F-statistic):	1.05e-264
Time:	16:54:36	Log-Likelihood:	-5865.8
No. Observations:	586	AIC:	1.180e+04
Df Residuals:	553	BIC:	1.194e+04
Df Model:	32		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
hourly_tweet_number	-14.7449	2.804	-5.258	0.000	-20.253	-9.237
hourly_retweet_number	-5.7622	2.417	-2.384	0.017	-10.509	-1.015
hourly_sum_of_followers	-0.0002	1.64e-05	-9.541	0.000	-0.000	-0.000
hourly_max_follower	0.0006	0.000	5.407	0.000	0.000	0.001
hourly_user_mentions	0.8510	0.707	1.204	0.229	-0.537	2.239
hourly_hashtags_in_tweet	2.9560	0.306	9.665	0.000	2.355	3.557
hourly_reply	-14.7449	2.804	-5.258	0.000	-20.253	-9.237
hourly_favorite_count	-1.7934	0.264	-6.783	0.000	-2.313	-1.274
hourly_verified	-2.563e-11	2.47e-10	-0.104	0.917	-5.1e-10	4.59e-10
hourly_number_of_citations	-1.9360	0.816	-2.372	0.018	-3.540	-0.333
hourly_ranking_scores	5.6654	1.173	4.829	0.000	3.361	7.970
time_of_day_0.0	-788.8048	1119.797	-0.704	0.481	-2988.381	1410.771
time_of_day_1.0	-512.5921	1111.048	-0.461	0.645	-2694.983	1669.799
time_of_day_2.0	-514.8152	1111.427	-0.463	0.643	-2697.951	1668.320
time_of_day_3.0	-861.1337	1116.786	-0.771	0.441	-3054.796	1332.528
time_of_day_4.0	-726.5447	1112.590	-0.653	0.514	-2911.964	1458.875
time_of_day_5.0	-835.5195	1124.138	-0.743	0.458	-3043.622	1372.583
time_of_day_6.0	-961.1354	1121.381	-0.857	0.392	-3163.822	1241.551
time_of_day_7.0	-1515.9941	1130.822	-1.341	0.181	-3737.226	705.238
time_of_day_8.0	-690.0694	1134.570	-0.608	0.543	-2918.664	1538.525
time_of_day_9.0	-1810.9710	1132.181	-1.600	0.110	-4034.873	412.931
time_of_day_10.0	-2125.9803	1162.432	-1.829	0.068	-4409.303	157.342
time_of_day_11.0	-1916.5693	1165.103	-1.645	0.101	-4205.139	372.000
time_of_day_12.0	-1845.4840	1158.630	-1.593	0.112	-4121.338	430.370
time_of_day_13.0	269.5574	1160.175	0.232	0.816	-2009.331	2548.446
time_of_day_14.0	3588.6745	1177.384	3.048	0.002	1275.983	5901.366
time_of_day_15.0	-3060.0826	1180.227	-2.593	0.010	-5378.358	-741.807
time_of_day_16.0	-1519.4036	1176.583	-1.291	0.197	-3830.522	791.715
time_of_day_17.0	-252.8866	1177.062	-0.215	0.830	-2564.946	2059.173
time_of_day_18.0	-1955.7685	1179.382	-1.658	0.098	-4272.384	360.847
time_of_day_19.0	-2367.9599	1169.489	-2.025	0.043	-4665.144	-70.775
time_of_day_20.0	-891.4442	1148.969	-0.776	0.438	-3148.321	1365.433
time_of_day_21.0	-1067.0133	1140.006	-0.936	0.350	-3306.285	1172.259
time_of_day_22.0	-1037.5571	1137.170	-0.912	0.362	-3271.257	1196.143
time_of_day_23.0	-1211.9980	1143.368	-1.060	0.290	-3457.874	1033.878

Omnibus:	940.828	Durbin-Watson:	1.858
Prob(Omnibus):	0.000	Jarque-Bera (JB):	844951.063
Skew:	8.920	Prob(JB):	0.00
Kurtosis:	188.168	Cond. No.	1.13e+21

Figure 15: OLS Regression Results for #SuperBowl

The RMSE is 5383 and the R-squared value is 0.909.

2.3.3 #SB49

OLS Regression Results

Dep. Variable:	target_value	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.897
Method:	Least Squares	F-statistic:	159.9
Date:	Fri, 09 Mar 2018	Prob (F-statistic):	2.88e-255
Time:	16:51:51	Log-Likelihood:	-5518.9
No. Observations:	583	AIC:	1.110e+04
Df Residuals:	550	BIC:	1.125e+04
Df Model:	32		
Covariance Type:	nonrobust		

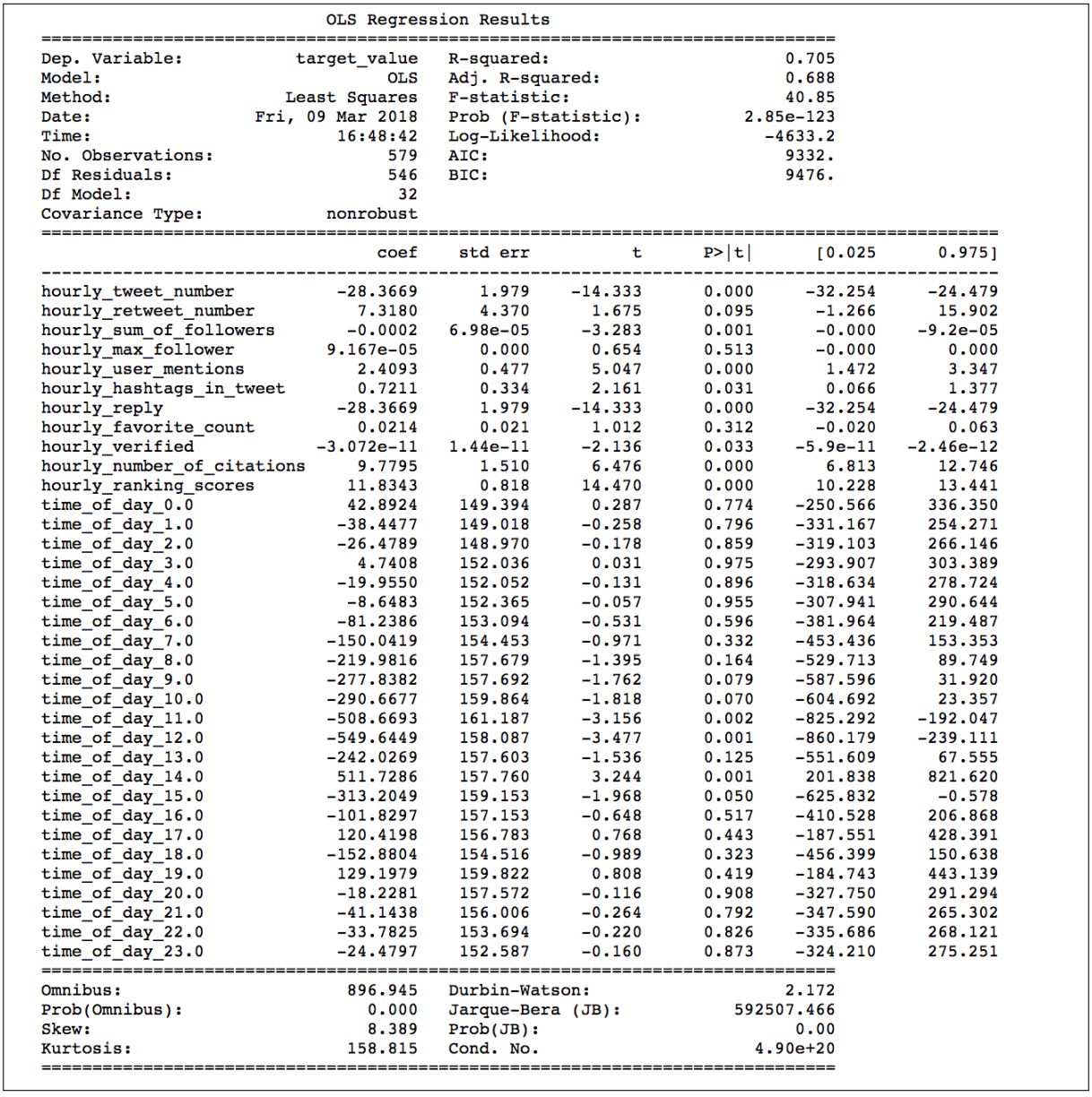
	coef	std err	t	P>|t|	[0.025	0.975]
hourly_tweet_number	-36.0404	3.621	-9.953	0.000	-43.153	-28.928
hourly_retweet_number	57.3253	18.537	3.092	0.002	20.912	93.738
hourly_sum_of_followers	0.0002	1.88e-05	10.917	0.000	0.000	0.000
hourly_max_follower	-0.0004	6.25e-05	-6.460	0.000	-0.001	-0.000
hourly_user_mentions	3.9544	1.114	3.550	0.000	1.766	6.142
hourly_hashtags_in_tweet	2.6687	0.541	4.932	0.000	1.606	3.731
hourly_reply	-36.0402	3.621	-9.953	0.000	-43.153	-28.927
hourly_favorite_count	-0.3168	0.091	-3.478	0.001	-0.496	-0.138
hourly_verified	-1.965e-08	3.37e-08	-0.583	0.560	-8.59e-08	4.66e-08
hourly_number_of_citations	3.9984	1.958	2.042	0.042	0.152	7.845
hourly_ranking_scores	13.5988	1.416	9.603	0.000	10.817	16.381
time_of_day_0.0	-181.6616	644.934	-0.282	0.778	-1448.498	1085.175
time_of_day_1.0	-35.7396	646.901	-0.055	0.956	-1306.439	1234.960
time_of_day_2.0	-823.9450	649.418	-1.269	0.205	-2099.588	451.698
time_of_day_3.0	-925.9397	651.040	-1.422	0.156	-2204.769	352.890
time_of_day_4.0	-864.5354	657.837	-1.314	0.189	-2156.716	427.645
time_of_day_5.0	1387.3038	655.786	2.115	0.035	99.152	2675.455
time_of_day_6.0	228.4420	665.692	0.343	0.732	-1079.168	1536.052
time_of_day_7.0	-1689.2993	670.921	-2.518	0.012	-3007.180	-371.418
time_of_day_8.0	-1329.9345	664.433	-2.002	0.046	-2635.071	-24.798
time_of_day_9.0	-990.1714	680.769	-1.454	0.146	-2327.398	347.055
time_of_day_10.0	-916.5785	670.659	-1.367	0.172	-2233.944	400.787
time_of_day_11.0	-84.8275	668.572	-0.127	0.899	-1398.095	1228.440
time_of_day_12.0	-538.3469	678.458	-0.793	0.428	-1871.033	794.339
time_of_day_13.0	-923.9268	665.990	-1.387	0.166	-2232.122	384.269
time_of_day_14.0	-330.2842	672.877	-0.491	0.624	-1652.007	991.439
time_of_day_15.0	181.7806	671.919	0.271	0.787	-1138.061	1501.622
time_of_day_16.0	297.1929	662.760	0.448	0.654	-1004.658	1599.044
time_of_day_17.0	396.7865	666.583	0.595	0.552	-912.574	1706.147
time_of_day_18.0	126.7615	659.066	0.192	0.848	-1167.833	1421.356
time_of_day_19.0	273.5388	659.678	0.415	0.679	-1022.258	1569.336
time_of_day_20.0	243.7370	658.299	0.370	0.711	-1049.350	1536.824
time_of_day_21.0	75.1390	657.237	0.114	0.909	-1215.862	1366.140
time_of_day_22.0	29.1049	657.749	0.044	0.965	-1262.902	1321.112
time_of_day_23.0	-104.7488	657.714	-0.159	0.874	-1396.688	1187.190

Omnibus:	951.785	Durbin-Watson:	1.515
Prob(Omnibus):	0.000	Jarque-Bera (JB):	677400.042
Skew:	9.361	Prob(JB):	0.00
Kurtosis:	168.938	Cond. No.	1.98e+21

Figure 16: OLS Regression Results for #SB49

The RMSE is 3126 and the R-squared value is 0.903, indicating that it is a very well-fitted model.

2.3.4 #GoHawks



OLS Regression Results

Dep. Variable:	target_value	R-squared:	0.705
Model:	OLS	Adj. R-squared:	0.688
Method:	Least Squares	F-statistic:	40.85
Date:	Fri, 09 Mar 2018	Prob (F-statistic):	2.85e-123
Time:	16:48:42	Log-Likelihood:	-4633.2
No. Observations:	579	AIC:	9332.
Df Residuals:	546	BIC:	9476.
Df Model:	32		
Covariance Type:	nonrobust		

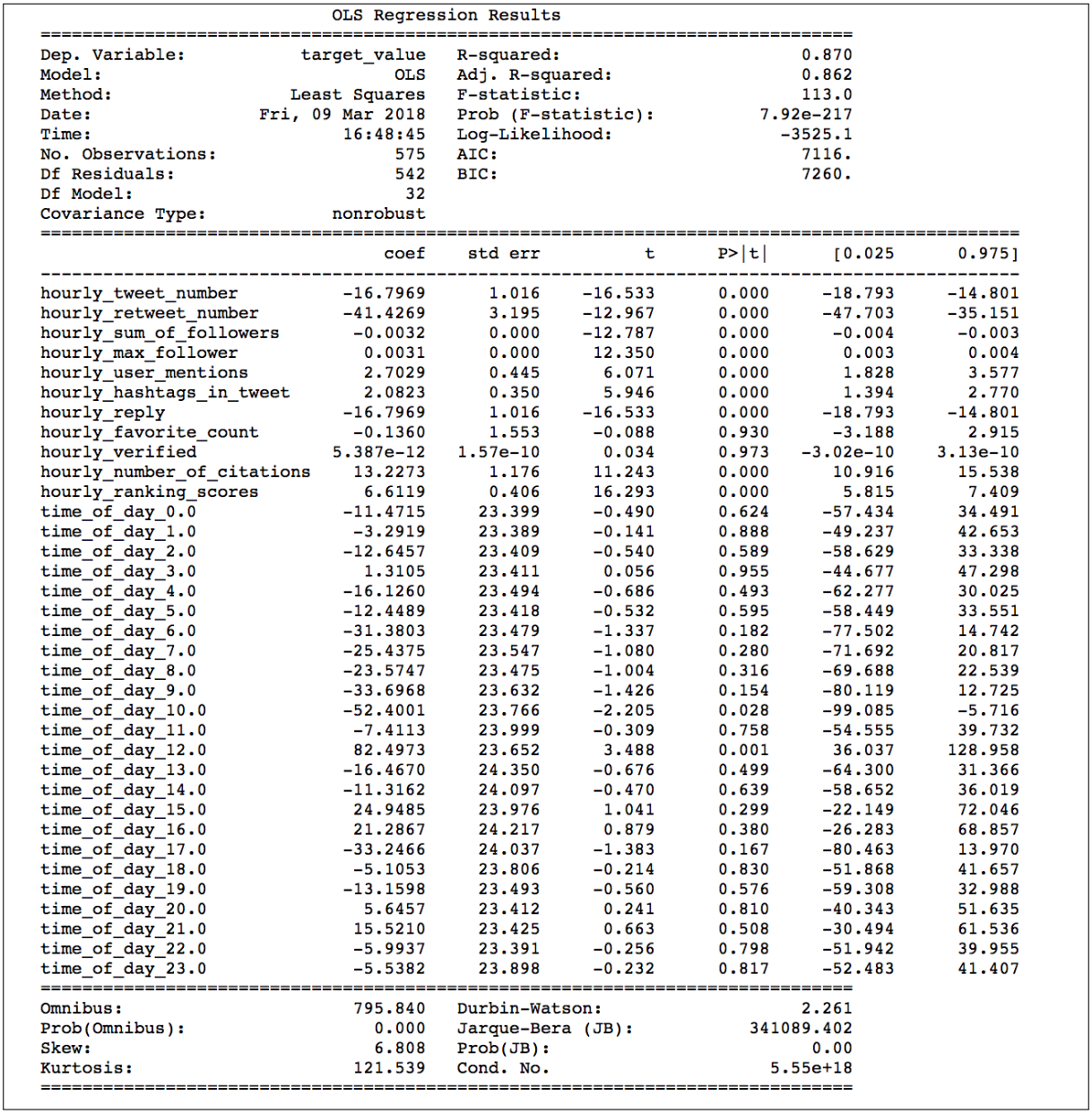
	coef	std err	t	P>|t|	[0.025	0.975]
hourly_tweet_number	-28.3669	1.979	-14.333	0.000	-32.254	-24.479
hourly_retweet_number	7.3180	4.370	1.675	0.095	-1.266	15.902
hourly_sum_of_followers	-0.0002	6.98e-05	-3.283	0.001	-0.000	-9.2e-05
hourly_max_follower	9.167e-05	0.000	0.654	0.513	-0.000	0.000
hourly_user_mentions	2.4093	0.477	5.047	0.000	1.472	3.347
hourly_hashtags_in_tweet	0.7211	0.334	2.161	0.031	0.066	1.377
hourly_reply	-28.3669	1.979	-14.333	0.000	-32.254	-24.479
hourly_favorite_count	0.0214	0.021	1.012	0.312	-0.020	0.063
hourly_verified	-3.072e-11	1.44e-11	-2.136	0.033	-5.9e-11	-2.46e-12
hourly_number_of_citations	9.7795	1.510	6.476	0.000	6.813	12.746
hourly_ranking_scores	11.8343	0.818	14.470	0.000	10.228	13.441
time_of_day_0.0	42.8924	149.394	0.287	0.774	-250.566	336.350
time_of_day_1.0	-38.4477	149.018	-0.258	0.796	-331.167	254.271
time_of_day_2.0	-26.4789	148.970	-0.178	0.859	-319.103	266.146
time_of_day_3.0	4.7408	152.036	0.031	0.975	-293.907	303.389
time_of_day_4.0	-19.9550	152.052	-0.131	0.896	-318.634	278.724
time_of_day_5.0	-8.6483	152.365	-0.057	0.955	-307.941	290.644
time_of_day_6.0	-81.2386	153.094	-0.531	0.596	-381.964	219.487
time_of_day_7.0	-150.0419	154.453	-0.971	0.332	-453.436	153.353
time_of_day_8.0	-219.9816	157.679	-1.395	0.164	-529.713	89.749
time_of_day_9.0	-277.8382	157.692	-1.762	0.079	-587.596	31.920
time_of_day_10.0	-290.6677	159.864	-1.818	0.070	-604.692	23.357
time_of_day_11.0	-508.6693	161.187	-3.156	0.002	-825.292	-192.047
time_of_day_12.0	-549.6449	158.087	-3.477	0.001	-860.179	-239.111
time_of_day_13.0	-242.0269	157.603	-1.536	0.125	-551.609	67.555
time_of_day_14.0	511.7286	157.760	3.244	0.001	201.838	821.620
time_of_day_15.0	-313.2049	159.153	-1.968	0.050	-625.832	-0.578
time_of_day_16.0	-101.8297	157.153	-0.648	0.517	-410.528	206.868
time_of_day_17.0	120.4198	156.783	0.768	0.443	-187.551	428.391
time_of_day_18.0	-152.8804	154.516	-0.989	0.323	-456.399	150.638
time_of_day_19.0	129.1979	159.822	0.808	0.419	-184.743	443.139
time_of_day_20.0	-18.2281	157.572	-0.116	0.908	-327.750	291.294
time_of_day_21.0	-41.1438	156.006	-0.264	0.792	-347.590	265.302
time_of_day_22.0	-33.7825	153.694	-0.220	0.826	-335.686	268.121
time_of_day_23.0	-24.4797	152.587	-0.160	0.873	-324.210	275.251

Omnibus:	896.945	Durbin-Watson:	2.172
Prob(Omnibus):	0.000	Jarque-Bera (JB):	592507.466
Skew:	8.389	Prob(JB):	0.00
Kurtosis:	158.815	Cond. No.	4.90e+20

Figure 17: OLS Regression Results for #GoHawks

The RMSE is 722.7 and the R-squared value is 0.705.

2.3.5 #GoPatriots



OLS Regression Results

Dep. Variable:	target_value	R-squared:	0.870
Model:	OLS	Adj. R-squared:	0.862
Method:	Least Squares	F-statistic:	113.0
Date:	Fri, 09 Mar 2018	Prob (F-statistic):	7.92e-217
Time:	16:48:45	Log-Likelihood:	-3525.1
No. Observations:	575	AIC:	7116.
Df Residuals:	542	BIC:	7260.
Df Model:	32		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
hourly_tweet_number	-16.7969	1.016	-16.533	0.000	-18.793	-14.801
hourly_retweet_number	-41.4269	3.195	-12.967	0.000	-47.703	-35.151
hourly_sum_of_followers	-0.0032	0.000	-12.787	0.000	-0.004	-0.003
hourly_max_follower	0.0031	0.000	12.350	0.000	0.003	0.004
hourly_user_mentions	2.7029	0.445	6.071	0.000	1.828	3.577
hourly_hashtags_in_tweet	2.0823	0.350	5.946	0.000	1.394	2.770
hourly_reply	-16.7969	1.016	-16.533	0.000	-18.793	-14.801
hourly_favorite_count	-0.1360	1.553	-0.088	0.930	-3.188	2.915
hourly_verified	5.387e-12	1.57e-10	0.034	0.973	-3.02e-10	3.13e-10
hourly_number_of_citations	13.2273	1.176	11.243	0.000	10.916	15.538
hourly_ranking_scores	6.6119	0.406	16.293	0.000	5.815	7.409
time_of_day_0.0	-11.4715	23.399	-0.490	0.624	-57.434	34.491
time_of_day_1.0	-3.2919	23.389	-0.141	0.888	-49.237	42.653
time_of_day_2.0	-12.6457	23.409	-0.540	0.589	-58.629	33.338
time_of_day_3.0	1.3105	23.411	0.056	0.955	-44.677	47.298
time_of_day_4.0	-16.1260	23.494	-0.686	0.493	-62.277	30.025
time_of_day_5.0	-12.4489	23.418	-0.532	0.595	-58.449	33.551
time_of_day_6.0	-31.3803	23.479	-1.337	0.182	-77.502	14.742
time_of_day_7.0	-25.4375	23.547	-1.080	0.280	-71.692	20.817
time_of_day_8.0	-23.5747	23.475	-1.004	0.316	-69.688	22.539
time_of_day_9.0	-33.6968	23.632	-1.426	0.154	-80.119	12.725
time_of_day_10.0	-52.4001	23.766	-2.205	0.028	-99.085	-5.716
time_of_day_11.0	-7.4113	23.999	-0.309	0.758	-54.555	39.732
time_of_day_12.0	82.4973	23.652	3.488	0.001	36.037	128.958
time_of_day_13.0	-16.4670	24.350	-0.676	0.499	-64.300	31.366
time_of_day_14.0	-11.3162	24.097	-0.470	0.639	-58.652	36.019
time_of_day_15.0	24.9485	23.976	1.041	0.299	-22.149	72.046
time_of_day_16.0	21.2867	24.217	0.879	0.380	-26.283	68.857
time_of_day_17.0	-33.2466	24.037	-1.383	0.167	-80.463	13.970
time_of_day_18.0	-5.1053	23.806	-0.214	0.830	-51.868	41.657
time_of_day_19.0	-13.1598	23.493	-0.560	0.576	-59.308	32.988
time_of_day_20.0	5.6457	23.412	0.241	0.810	-40.343	51.635
time_of_day_21.0	15.5210	23.425	0.663	0.508	-30.494	61.536
time_of_day_22.0	-5.9937	23.391	-0.256	0.798	-51.942	39.955
time_of_day_23.0	-5.5382	23.898	-0.232	0.817	-52.483	41.407

Omnibus:	795.840	Durbin-Watson:	2.261
Prob(Omnibus):	0.000	Jarque-Bera (JB):	341089.402
Skew:	6.808	Prob(JB):	0.00
Kurtosis:	121.539	Cond. No.	5.55e+18

Figure 18: OLS Regression Results for #GoPatriots

The RMSE is 111.2 and the R-squared value is 0.870, indicating a model that is well-fitted.

2.3.6 #Patriots

OLS Regression Results

Dep. Variable:	target_value	R-squared:	0.822
Model:	OLS	Adj. R-squared:	0.812
Method:	Least Squares	F-statistic:	79.86
Date:	Fri, 09 Mar 2018	Prob (F-statistic):	7.99e-185
Time:	16:50:17	Log-Likelihood:	-5247.4
No. Observations:	587	AIC:	1.056e+04
Df Residuals:	554	BIC:	1.071e+04
Df Model:	32		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
hourly_tweet_number	-30.7896	2.481	-12.409	0.000	-35.663	-25.916
hourly_retweet_number	-30.0885	3.349	-8.984	0.000	-36.667	-23.510
hourly_sum_of_followers	0.0003	4.94e-05	6.831	0.000	0.000	0.000
hourly_max_follower	-0.0005	0.000	-5.207	0.000	-0.001	-0.000
hourly_user_mentions	6.8760	0.955	7.203	0.000	5.001	8.751
hourly_hashtags_in_tweet	3.8130	0.407	9.376	0.000	3.014	4.612
hourly_reply	-30.7895	2.481	-12.409	0.000	-35.663	-25.916
hourly_favorite_count	0.2281	0.173	1.317	0.188	-0.112	0.568
hourly_verified	-2.317e-10	8.87e-11	-2.613	0.009	-4.06e-10	-5.75e-11
hourly_number_of_citations	-4.3672	1.711	-2.553	0.011	-7.728	-1.006
hourly_ranking_scores	11.0860	0.951	11.651	0.000	9.217	12.955
time_of_day_0.0	148.9641	381.385	0.391	0.696	-600.173	898.101
time_of_day_1.0	-111.7876	382.380	-0.292	0.770	-862.881	639.305
time_of_day_2.0	-60.3311	381.378	-0.158	0.874	-809.454	688.792
time_of_day_3.0	-161.9209	381.159	-0.425	0.671	-910.614	586.772
time_of_day_4.0	-137.1761	382.492	-0.359	0.720	-888.488	614.136
time_of_day_5.0	-182.4326	384.723	-0.474	0.636	-938.127	573.262
time_of_day_6.0	-373.2976	388.789	-0.960	0.337	-1136.978	390.383
time_of_day_7.0	-706.7405	392.844	-1.799	0.073	-1478.386	64.905
time_of_day_8.0	-826.6608	394.045	-2.098	0.036	-1600.667	-52.655
time_of_day_9.0	-426.1855	402.504	-1.059	0.290	-1216.805	364.434
time_of_day_10.0	921.0734	392.831	2.345	0.019	149.454	1692.693
time_of_day_11.0	-891.6236	402.133	-2.217	0.027	-1681.516	-101.731
time_of_day_12.0	-769.8412	403.476	-1.908	0.057	-1562.371	22.688
time_of_day_13.0	-604.7208	404.768	-1.494	0.136	-1399.788	190.347
time_of_day_14.0	-352.8807	408.239	-0.864	0.388	-1154.767	449.005
time_of_day_15.0	-618.6296	405.590	-1.525	0.128	-1415.312	178.053
time_of_day_16.0	-174.4470	402.720	-0.433	0.665	-965.492	616.598
time_of_day_17.0	76.3784	401.497	0.190	0.849	-712.264	865.021
time_of_day_18.0	196.4644	403.821	0.487	0.627	-596.743	989.672
time_of_day_19.0	-335.6422	401.147	-0.837	0.403	-1123.597	452.313
time_of_day_20.0	270.2140	402.889	0.671	0.503	-521.163	1061.591
time_of_day_21.0	-464.4863	399.227	-1.163	0.245	-1248.670	319.697
time_of_day_22.0	80.0097	393.181	0.203	0.839	-692.298	852.317
time_of_day_23.0	142.5093	391.632	0.364	0.716	-626.756	911.774

Omnibus:	1044.785	Durbin-Watson:	1.884
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1056563.374
Skew:	11.134	Prob(JB):	0.00
Kurtosis:	209.646	Cond. No.	1.71e+20

Figure 19: OLS Regression Results for #Patriots

The RMSE is 1853 and the R-squared value is 0.822, which suggests that the model is well-fitted.

2.3.7 Analysis

The fitting accuracy and R-squared values are compared to the model used in the previous section.

		#NFL	#SuperBowl	#SB49	#GoHawks	#GoPatriots	#Patriots
5-Var	RMSE	567.1	6443	3800	911	188.2	2302
	R^2	0.585	0.870	0.857	0.532	0.627	0.723
12-Var	RMSE	444.9	5383	3126	722.7	111.2	1853
	R^2	0.745	0.909	0.903	0.705	0.870	0.822

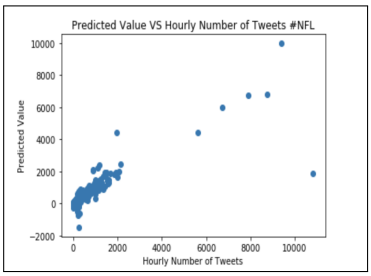
Table 2: RMSE and R-Squared Values for 5-Variable and 12-Variable Models

The addition of the 7 features has improved the accuracy of the linear regression model. This is due to some of the additional features having greater significance than the features used in the initial model, which can be confirmed using the OLS regression results.

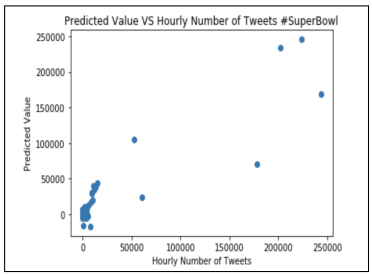
The top 3 features across all hashtags are (1) Hourly number of tweets, (2) Hourly number of replies, and (3) Ranking score.

They are selected from the most significant features for each hashtag, which are chosen according to the p-values.

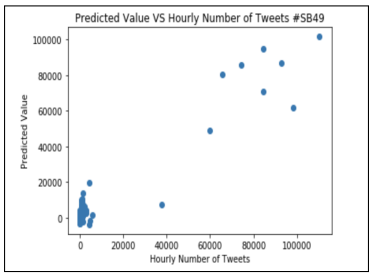
For each of the top 3 features, a scatter plot of predictant against the value of the feature is shown below.



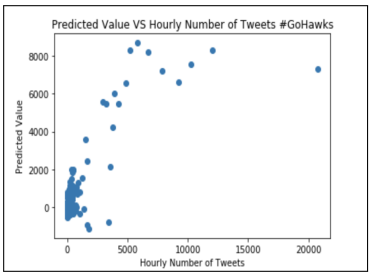
(a) #NFL



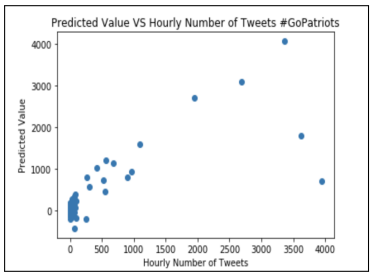
(b) #SuperBowl



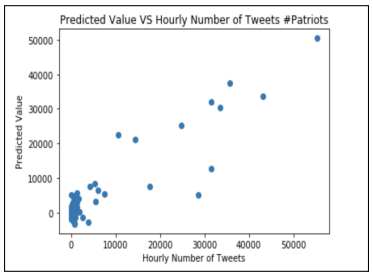
(c) #SB49



(d) #GoHawks

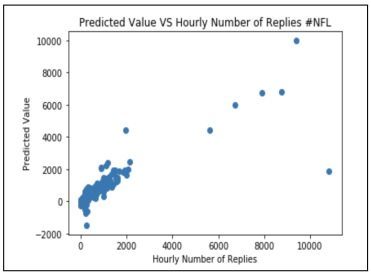


(e) #GoPatriots

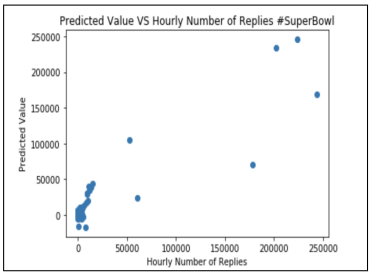


(f) #Patriots

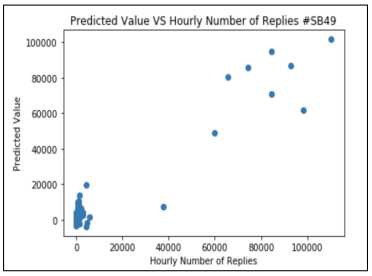
Figure 20: Predictant Against Hourly Number of Tweets



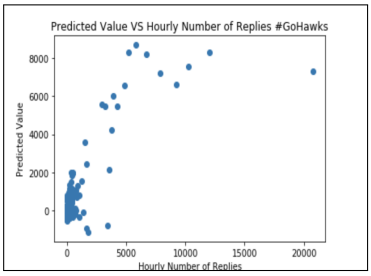
(a) #NFL



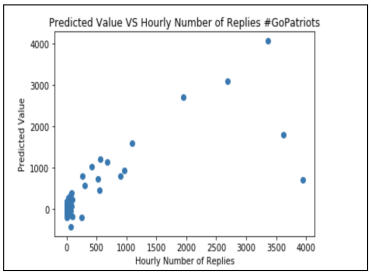
(b) #SuperBowl



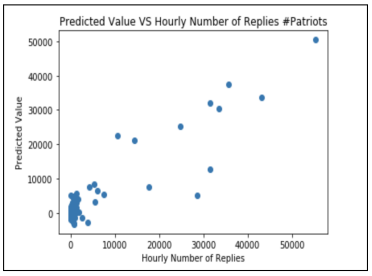
(c) #SB49



(d) #GoHawks

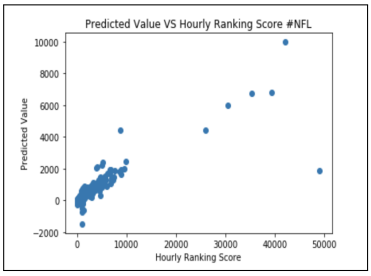


(e) #GoPatriots

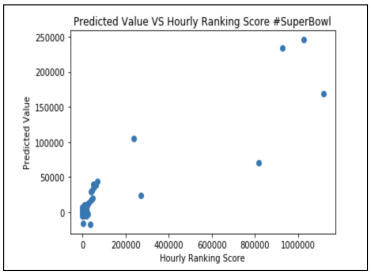


(f) #Patriots

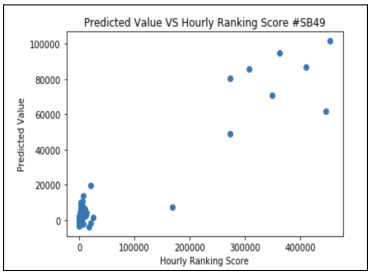
Figure 21: Predictant Against Hourly Number of Replies



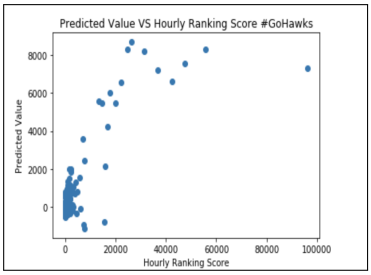
(a) #NFL



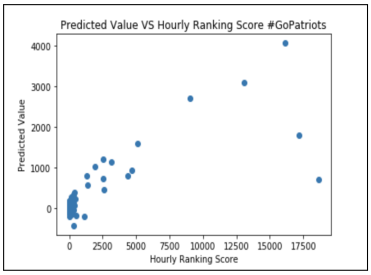
(b) #SuperBowl



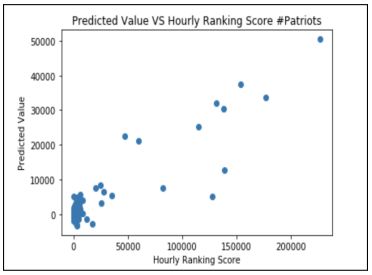
(c) #SB49



(d) #GoHawks



(e) #GoPatriots



(f) #Patriots

Figure 22: Predictant Against Ranking Score

From the plots, a relatively linear relationship can be observed between the top features and the target value, suggesting that the features were well-designed. Furthermore, the plots across features for the same hashtag are similarly shaped, hence the top features also have approximately the same significance.

2.4 k-Fold Cross-Validation

In this section, different regression models are used to predict the number of tweets for three time periods:

1. Before Feb. 1, 8:00 a.m.
2. Between Feb. 1, 8:00 a.m. and 8:00 p.m.
3. After Feb. 1, 8:00 p.m.

k-fold cross-validation is used to evaluate the performance of the prediction models. The average mean absolute error (MAE) and average RMSE are computed and presented for each evaluation.

2.4.1 Separate Hashtags

Three types of regression models are used to predict the number of tweets for each hashtag over the three time periods previously mentioned. The regression models used are:

- Linear Regression
- Support Vector Regression
- Random Forest Regression

2.4.1.1 #NFL

Time Period	Regression Model	Average MAE	Average RMSE
Before Feb. 1, 8:00 a.m.	Linear	84.95	9.217
	Support Vector	165.3	12.86
	Random Forest	70.31	8.385
Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Linear	17030	130.5
	Support Vector	5221	72.25
	Random Forest	1912	43.73
After Feb. 1, 8:00 p.m.	Linear	109.7	10.48
	Support Vector	294.8	17.17
	Random Forest	135.4	11.64
Entire	Linear	122.5	11.07
	Support Vector	264.6	16.27
	Random Forest	133.4	11.55

Table 3: Errors for Different Time Periods

2.4.1.2 #SuperBowl

Time Period	Regression Model	Average MAE	Average RMSE
Before Feb. 1, 8:00 a.m.	Linear	177.8	13.34
	Support Vector	249.4	15.79
	Random Forest	157.3	12.54
Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Linear	96570	310.8
	Support Vector	120100	346.6
	Random Forest	51300	226.5
After Feb. 1, 8:00 p.m.	Linear	278.7	16.70
	Support Vector	585.4	24.20
	Random Forest	276.9	16.64
Entire	Linear	1154	33.97
	Support Vector	1392	37.31
	Random Forest	1253	35.40

Table 4: Errors for Different Time Periods

2.4.1.3 #SB49

Time Period	Regression Model	Average MAE	Average RMSE
Before Feb. 1, 8:00 a.m.	Linear	48.10	6.936
	Support Vector	103.5	10.17
	Random Forest	49.88	7.063
Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Linear	97850	312.8
	Support Vector	31520	177.5
	Random Forest	28986	170.3
After Feb. 1, 8:00 p.m.	Linear	119.7	10.94
	Support Vector	344.7	18.57
	Random Forest	122.6	11.07
Entire	Linear	814.1	28.53
	Support Vector	1420	37.69
	Random Forest	1322	36.36

Table 5: Errors for Different Time Periods

2.4.1.4 #GoHawks

Time Period	Regression Model	Average MAE	Average RMSE
Before Feb. 1, 8:00 a.m.	Linear	198.4	14.09
	Support Vector	136.2	11.67
	Random Forest	74.72	8.644
Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Linear	43950	209.7
	Support Vector	3345	57.83
	Random Forest	2768	52.62
After Feb. 1, 8:00 p.m.	Linear	4563	67.55
	Support Vector	34.72	5.892
	Random Forest	23.08	4.805
Entire	Linear	211.7	14.55
	Support Vector	190.8	13.81
	Random Forest	101.3	10.06

Table 6: Errors for Different Time Periods

2.4.1.5 #GoPatriots

Time Period	Regression Model	Average MAE	Average RMSE
Before Feb. 1, 8:00 a.m.	Linear	15.34	3.916
	Support Vector	10.68	3.268
	Random Forest	7.998	2.828
Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Linear	1877	43.33
	Support Vector	1460	38.22
	Random Forest	895.9	29.93
After Feb. 1, 8:00 p.m.	Linear	2.567	1.602
	Support Vector	4.848	2.202
	Random Forest	3.366	1.835
Entire	Linear	47.69	6.905
	Support Vector	37.34	6.111
	Random Forest	32.10	5.666

Table 7: Errors for Different Time Periods

2.4.1.6 #Patriots

Time Period	Regression Model	Average MAE	Average RMSE
Before Feb. 1, 8:00 a.m.	Linear	164.9	12.84
	Support Vector	168.4	12.98
	Random Forest	102.0	10.10
Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Linear	59210	243.3
	Support Vector	12100	110.0
	Random Forest	15210	123.3
After Feb. 1, 8:00 p.m.	Linear	87.26	9.341
	Support Vector	142.3	11.93
	Random Forest	102.8	10.14
Entire	Linear	642.9	25.35
	Support Vector	483.8	22.00
	Random Forest	399.6	19.99

Table 8: Errors for Different Time Periods

2.4.1.7 Analysis

The average MAE and average RMSE during the time period between Feb. 1, 8:00 a.m. and 8:00 p.m. are much higher than those of time periods outside it. This could be due to the much higher number of tweets over a shorter period of time, leading to the model predicting less accurately. When predicting the number of tweets for separate time periods, the **random forest** regression model outperformed the other two models, which suggests that some features have a non-linear relationship with the number of tweets.

2.4.2 Aggregated Data

In this section, the data of all hashtags is aggregated and separate models are trained to predict the hourly number of tweets for each of the 3 time periods.

Time Period	Regression Model	Average MAE	Average RMSE
Before Feb. 1, 8:00 a.m.	Linear	366.2	19.14
	Support Vector	178.1	13.35
	Random Forest	120.6	10.98
Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Linear	36850	192.0
	Support Vector	84780	291.2
	Random Forest	39048	197.6
After Feb. 1, 8:00 p.m.	Linear	694.4	26.35
	Support Vector	1097	33.12
	Random Forest	349.0	18.68
Entire	Linear	4556	67.50
	Support Vector	1955	44.21
	Random Forest	1770	42.07

Table 9: Errors for Different Time Periods

The random forest regression models produced the lowest average MAE and average RMSE for all time periods, except for between Feb. 1, 8:00 a.m. and 8:00 p.m., where the linear regression model slightly outperforms it.

2.4.3 Comparison

Random forest regression generally yielded the best accuracy. In this section, the combined model using aggregated data is compared against the models used to predict the hourly number of tweets for separate hashtag. The average MAE for the different models are tabulated below.

Time Period	[1]	[2]	[3]	[4]	[5]	[6]	Combined
Before Feb. 1, 8:00 a.m.	70.31	157.3	49.88	74.72	7.998	102.0	120.6
Between Feb. 1, 8:00 a.m. and 8:00 p.m.	1912	51300	119.7	2768	895.9	15210	39048
After Feb. 1, 8:00 p.m.	135.4	276.9	122.6	23.08	3.366	102.8	349.0
Entire	133.4	1253	1322	101.3	32.10	399.6	1770

[1] #NFL
[2] #SuperBowl
[3] #SB49
[4] #GoHawks
[6] #GoPatriots
[6] #Patriots

Table 10: Average MAE for Separate and Combined Models (Random Forest Regression)

The separate models outperformed the combined models in all cases, except for the before Feb. 1, 8:00 a.m and between Feb. 1, 8:00 a.m. and 8:00 p.m. time periods for the #SuperBowl hashtag. The anomaly could be due to the large number of tweets within a shorter time period, similar to errors obtained in the previous sections. The better performance of separate models is expected because each hashtag are used by different groups of users and hence the significance of their respective features are also distinct.

2.5 Prediction for Other Hashtags

In this section, the number of tweets in the next hour for other hashtags is predicted using the best model determined in the previous section. A suitable hour is chosen from the test data set to evaluate the performance of the model. The actual and predicted number of tweets for each test file is tabulated below. The predicted values are fairly close to the actual values.

File Name	Actual	Predicted
sample1_period1	178	164
sample2_period2	100,221	82,923
sample3_period3	523	571
sample4_period1	201	213
sample5_period1	213	223
sample6_period2	37,307	32,305
sample7_period3	120	49
sample8_period1	11	26
sample9_period2	2790	2427
sample10_period3	61	53

Table 11: Actual and Predicted Number of Tweet in the Next Hour

3 Fan Base Prediction

In this section, binary classifiers are used to predict the location of the author of a tweet (Washington or Massachusetts) using the textual content of the tweet. Six types of classifiers were explored and for each model, the ROC curve, confusion matrix, accuracy, recall and precision are reported.

3.1 Support Vector Machine (SVM)

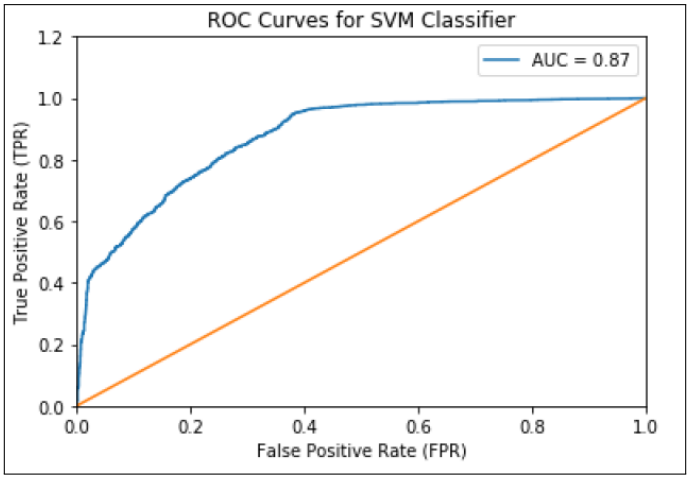


Figure 23: ROC Curve

		Predicted	
		Washington	Massachusetts
Actual	Washington	1143	726
	Massachusetts	100	2091

Table 12: Confusion Matrix

The accuracy is 0.7966, the recall is 0.7830, and the precision is 0.8309.

3.2 AdaBoost

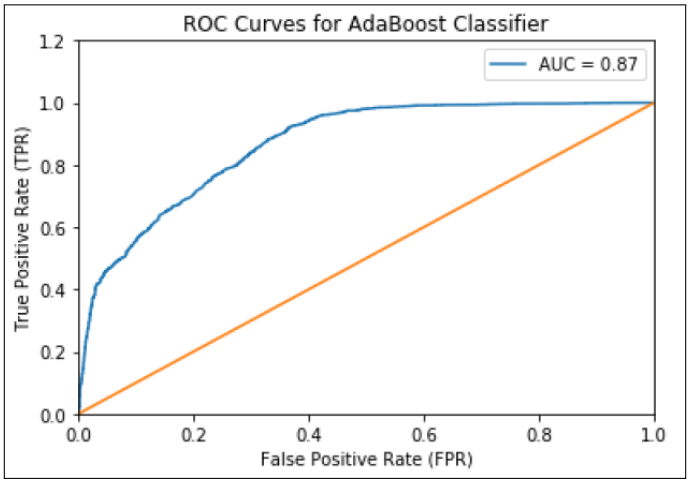


Figure 24: ROC Curve

		Predicted	
		Washington	Massachusetts
Actual	Washington	1236	633
	Massachusetts	243	1948

Table 13: Confusion Matrix

The accuracy is 0.7842, the recall is 0.7752, and the precision is 0.7952.

3.3 Random Forest

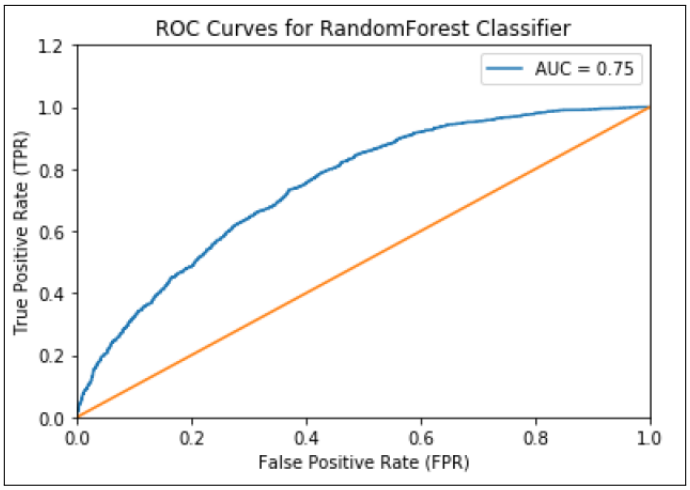


Figure 25: ROC Curve

		Predicted	
		Washington	Massachusetts
Actual	Washington	687	1182
	Massachusetts	143	2048

Table 14: Confusion Matrix

The accuracy is 0.6736, the recall is 0.6512, and the precision is 0.7309.

3.4 Neural Network (NN)

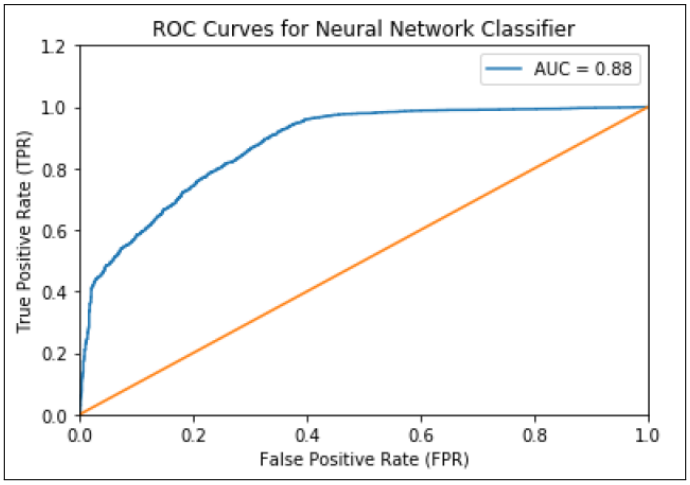


Figure 26: ROC Curve

		Predicted	
		Washington	Massachusetts
Actual	Washington	1153	713
	Massachusetts	115	2076

Table 15: Confusion Matrix

The accuracy is 0.7961, the recall is 0.7830, and the precision is 0.8269.

3.5 k-Nearest Neighbors (k-NN)

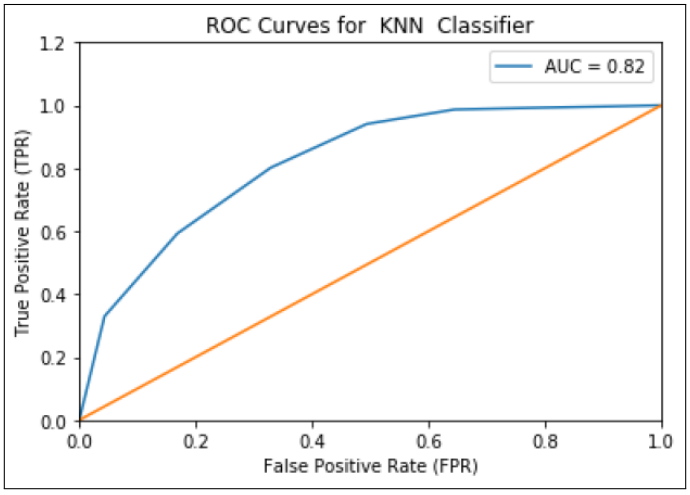


Figure 27: ROC Curve

		Predicted	
		Washington	Massachusetts
Actual	Washington	1254	615
	Massachusetts	434	1757

Table 16: Confusion Matrix

The accuracy is 0.7951, the recall is 0.7830, and the precision is 0.8203.

3.6 Gradient Boosting

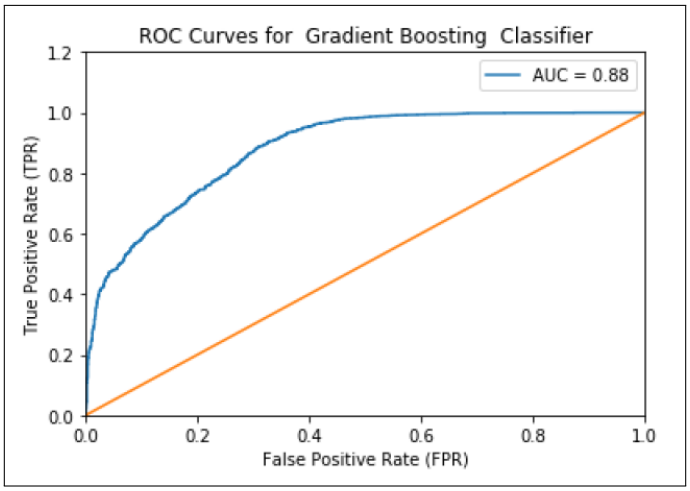


Figure 28: ROC Curve

		Predicted	
		Washington	Massachusetts
Actual	Washington	1179	690
	Massachusetts	142	2049

Table 17: Confusion Matrix

The accuracy is 0.7416, the recall is 0.7364, and the precision is 0.7418.

3.7 Analysis

	SVM	**AdaBoost**	**Random Forest**	**NN**	***k*-NN**	**Gradient Boosting**
Accuracy	0.7966	0.7842	0.6736	0.7961	0.7951	0.7416
Recall	0.7830	0.7752	0.6512	0.7830	0.7830	0.7364
Precision	0.8309	0.7952	0.7309	0.8269	0.8203	0.7418

Table 18: Accuracy, Recall and Precision for Different Classifiers

The SVM model produced the best results, with the highest accuracy, recall and precision among all six models.

4 Sentiment Prediction, Sentiment Comparison and Brand Sentiment

4.1 Background

Sentiment analysis is the process of identifying whether a given text carries a positive, neutral or negative opinion. It is an important tool which can provide machine learning engineers and data scientists key information about how a population feels about the topic of interest and narrow down texts which are either extremely negative or extremely positive to find out what went well and what can be done better. The use cases for sentiment analysis is endless whether it be companies looking to find out what they have done well and what can be improved further, or analysis on how a population feels about a sensitive subject.

4.2 Idea and Motivation

The initial idea was to predict the sentiment of a user's next tweet based on previous tweets by the same user. It was motivated by problems 1.2 and 1.3 of the project but to take a more personalized approach and see how well the linear prediction algorithm would work. Having analyzed the dataset, it was found that most users who are part of any hashtag have tweeted once and only about 100 users have tweeted more than three times. This limited number of data points with regards to the number of tweets for individual users would make it unsuitable for a linear regression problem.

Due to this limitation of the number of tweets by each user, the idea of predicting the sentiment of the next hours tweets based on the sentiment of the previous hours tweets arose as an idea which would be worth pursuing for this part. Another part of this proposed project is getting a sense of the number of positive and negative sentiment tweets which are posted over the span of the training data that has been provided and gain some insights about which team performed well during certain parts of the Super Bowl. Finally, another idea which has been implemented is sentiment analysis on advertisements by certain companies which were picked out from doing some research on companies who had good advertisements during the 2015 Super Bowl.

4.3 Setup

Since there are three different components which have been implemented in this section, the setup for each of the ideas will be covered in this part.

The first component which is predicting the sentiments of tweets for the next hour based on the sentiment of the last hour's tweet was done by first gathering all of the tweets for each hashtag. Each of the tweets was then parsed through the `TextBlob` package in Python and `TextBlob` returns a number between 1 and -1, with 1 being as positive as possible and -1 being as negative as possible. The value of 0 indicates a tweet which is neutral. The sentiment results are then bucketed into hours and the average sentiment for each hour is computed. Linear regression is carried out on this data with other features such as number of hashtags, whether the tweet was a retweet, the number of followers etc.

The second component involves looking at the number of positive and negative tweets by the hour and also getting an average sentiment for each hour. This was done by taking all of the tweets and finding the sentiment using `TextBlob`. Based on the score received, the positive or negative ratings count is increased by one. In addition to this, positive and negative tweets are categorized into being extremely positive, mildly positive, mildly negative and extremely negative. The reason for this is that most tweets are very close to being neutral but have a couple of words in there to make the sentiment analyzer categorize it is as being positive or negative. With this categorization, there is another level of granularity that has been added to the dataset and some more in-depth analysis can be performed. The boundaries for being mildly positive, extremely positive, mildly negative and extremely negative are [0, 0.3], [0.31, 1], [-0.01, -0.3] and [-0.31, -1] respectively. Since the maximum granularity with which `TextBlob` returns sentiment is in the hundredths, setting the boundaries in this manner makes it appropriate to categorize the data.

Since the dataset is spread over around 700 hours, the data has been analyzed and fitted for all 700 hours, a 5 day period around the Super Bowl including Super Bowl day, and a 17-hour period which spans 5 hours before and after the Super Bowl.

Finally, the third component implemented in this part of the project is obtaining sentiment for advertisements which were shown during the Super Bowl. In order to get the sentiments from users, tweets were collected from the beginning of the Super Bowl and 5 hours after it ended. The 5 hour period is added to take into account users who were unable to watch the commercials live during the Super Bowl.

4.4 Results and Analysis

In this section, the results for each of the three components are presented and analyzed.

4.4.1 Sentiment Prediction

Below are the scatter plots for the predicted against actual sentiment over time periods: all training data, 5 days including the day of the Super Bowl, and 17 hours which spans 5 hours before and after the Super Bowl.

4.4.1.1 Entire Period

#GoPatriots

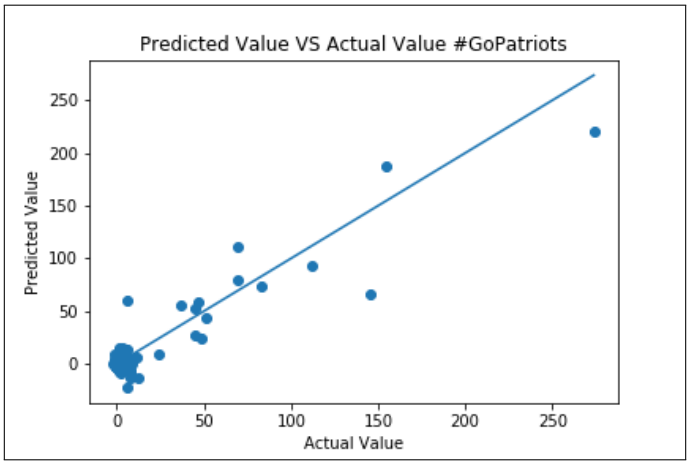


Figure 29: Predicted Sum of Sentiment Against Actual Sum of Sentiment

The average RMSE is 6.25 and the R-squared value is 0.859. It can be concluded that the predicted sentiment for #GoPatriots from one hour to the next matches the actual sentiment very well.

#GoHawks

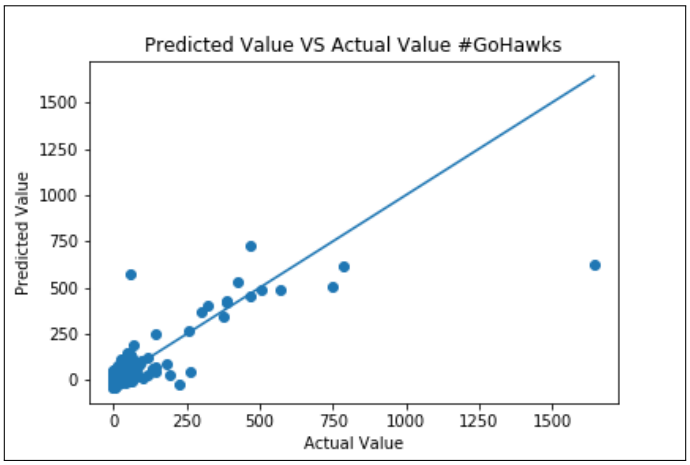


Figure 30: Predicted Sum of Sentiment Against Actual Sum of Sentiment

The average RMSE is 57.95 and the R-squared value is 0.666. The predicted sentiment for #GoHawks from one hour to the next matches the actual sentiment fairly well but not as good as #GoPatriots.

#NFL

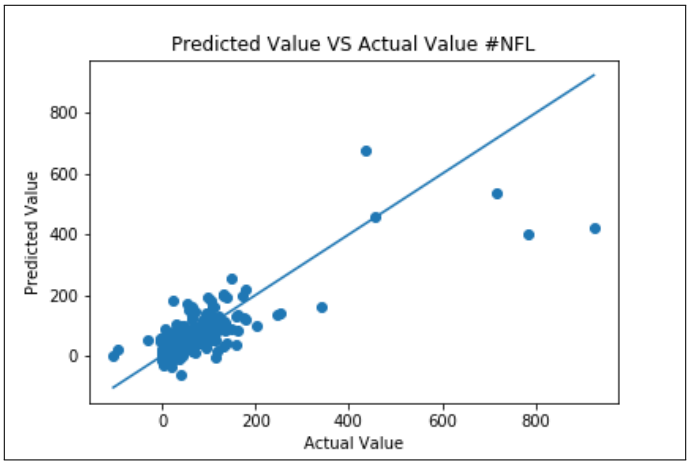


Figure 31: Predicted Sum of Sentiment Against Actual Sum of Sentiment

The average RMSE is 41.67 and the R-squared value is 0.671, which is slightly better than for #GoHawks. This is an interesting finding because #NFL contains tweets from a variety of topics which may not be related to the Super Bowl, unlike #GoHawks and #GoPatriots which are for team-specific.

#Patriots

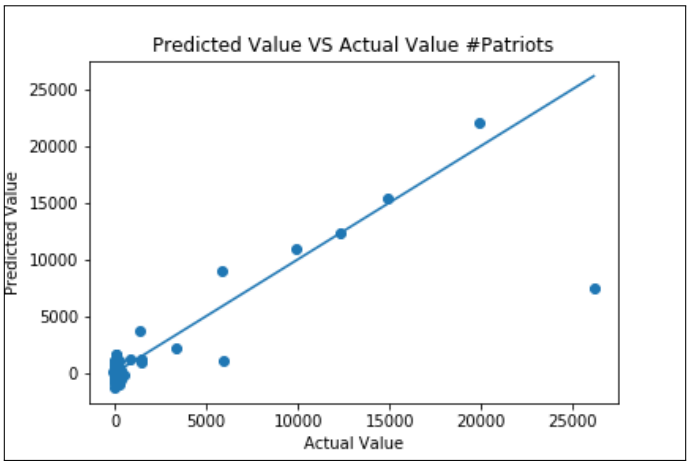


Figure 32: Predicted Sum of Sentiment Against Actual Sum of Sentiment

The average RMSE is 874.94 and the R-squared value is 0.727. From the results obtained, it can be concluded that the predicted sentiment matches the actual sentiment fairly well but it is worse than that of #GoPatriots. This can be attributed to a wide variety of reasons. During the period the data was collected, the Pro Bowl, which is on the weekend before the Super Bowl, was held as well. Patriots players who played the Pro Bowl could be the reason why there are so many more tweets for #Patriots as compared to #GoPatriots. The RMSE value for #Patriots is very high compared to the others that have been examined so far but that can be attributed to the much larger number of tweets which include #Patriots.

#SB49

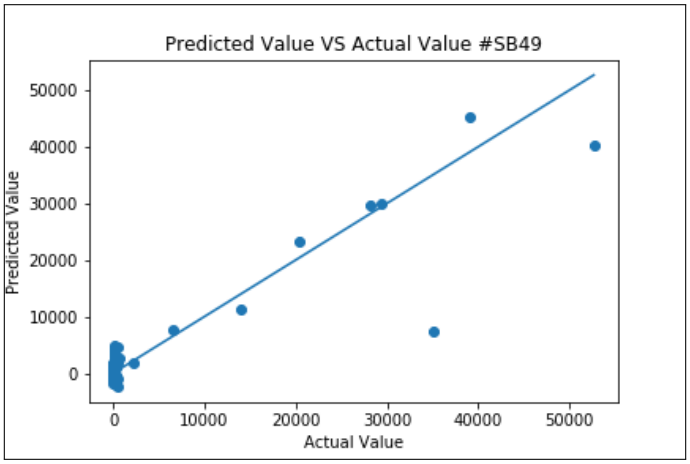


Figure 33: Predicted Sum of Sentiment Against Actual Sum of Sentiment

The average RMSE is 1480 and the R-squared value is 0.835. The large R-squared value could be due to a much larger dataset.

#SuperBowl

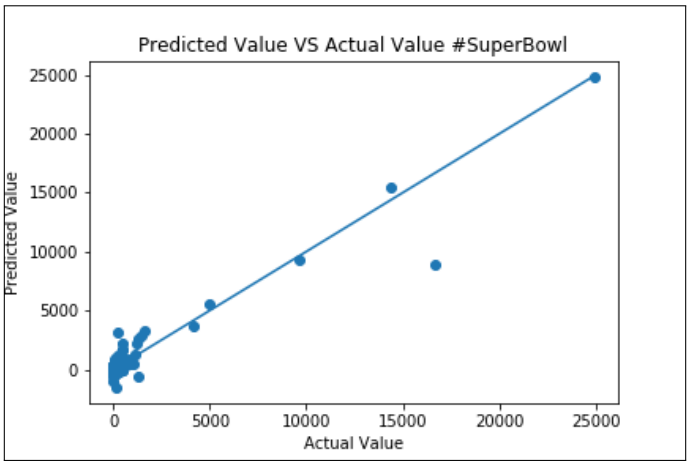


Figure 34: Predicted Sum of Sentiment Against Actual Sum of Sentiment

The average RMSE is 435.35 and the R-squared value is 0.910. This is the best R-squared obtained and it is an interesting finding as the hashtag can be used at any time and it does not pertain to supporters of any teams. Users also use it when watching the national anthem and the halftime show.

The average RMSE and R-squared value for each hashtag is compiled in the table below.

Hashtag	**Average RMSE**	**R-Squared Value**
#GoPatriots	6.25	0.859
#GoHawks	57.95	0.666
#NFL	41.67	0.671
#Patriots	874.9	0.727
#SB49	1480	0.835
#SuperBowl	435.4	0.91

Table 19: Sentiment Predictor Performance for Different Hashtags

4.4.1.2 5-Day Period

In this section, only the data within the 5-day period previously defined is considered.

#GoPatriots

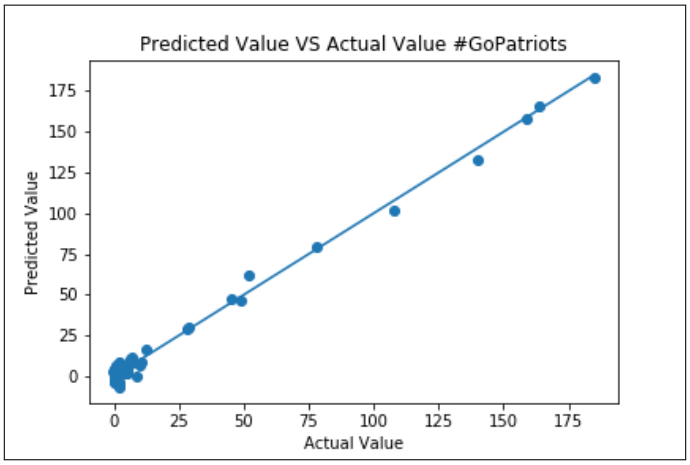


Figure 35: Predicted Sum of Sentiment Against Actual Sum of Sentiment

#GoHawks

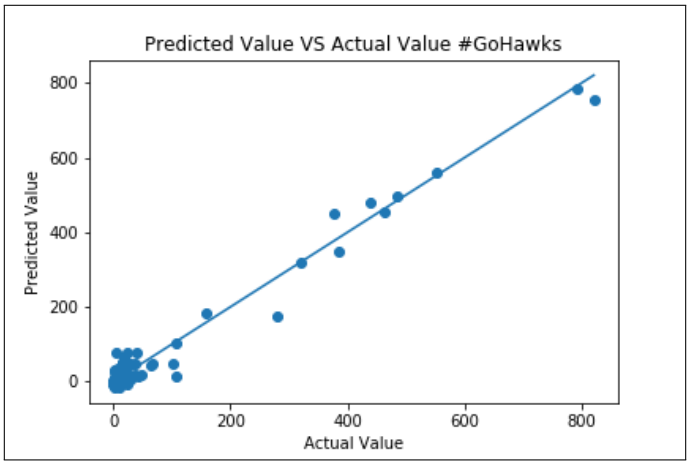


Figure 36: Predicted Sum of Sentiment Against Actual Sum of Sentiment

#NFL

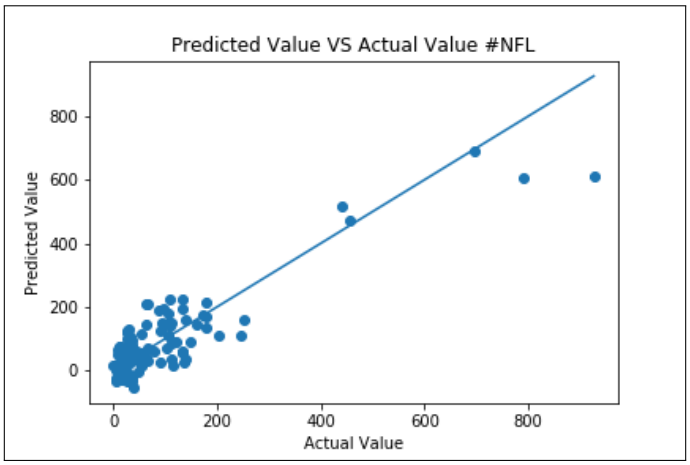


Figure 37: Predicted Sum of Sentiment Against Actual Sum of Sentiment

#Patriots

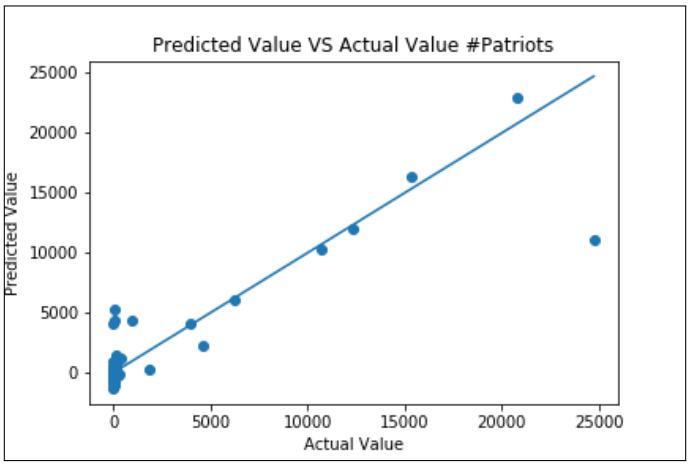


Figure 38: Predicted Sum of Sentiment Against Actual Sum of Sentiment

#SB49

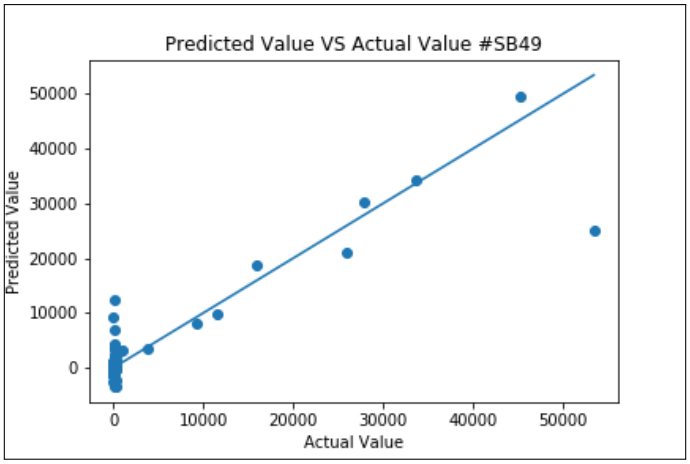


Figure 39: Predicted Sum of Sentiment Against Actual Sum of Sentiment

#SuperBowl

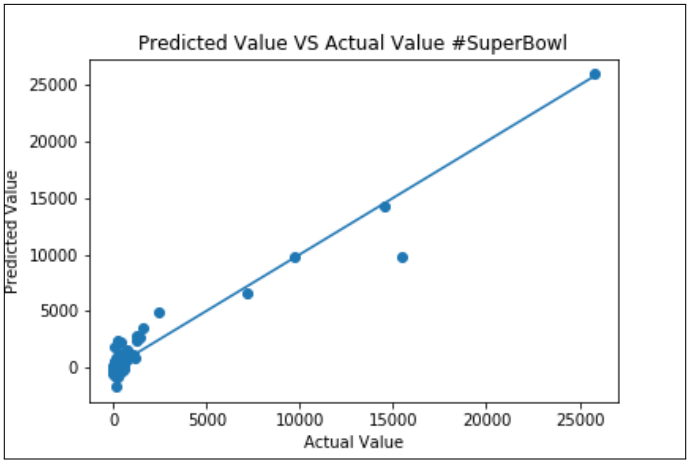


Figure 40: Predicted Sum of Sentiment Against Actual Sum of Sentiment

Hashtag	**Average RMSE**	***R*-Squared Value**
#GoPatriots	3.5	0.991
#GoHawks	25.8	0.973
#NFL	69.7	0.783
#Patriots	1757	0.804
#SB49	3730	0.819
#SuperBowl	893	0.935

Table 20: Sentiment Predictor Performance for Different Hashtags

The R-squared values for all of the hashtags increased, except for #SB49. This is expected because the time period has been narrowed down. For #SB49, the drop in R-squared value and increase in the average RMSE could be due to the smaller dataset as it is a more general hashtag. Upon further inspection, the

increase in the R-squared values for #SuperBowl and #NFL are not significant, suggesting that sentiments for generic hashtags do not change substantially when comparing between the whole training period and the 5-day period.

4.4.1.3 17-Hour Period

Finally, the dataset is trimmed to consider only the 17-hour period mentioned previously.

#GoPatriots

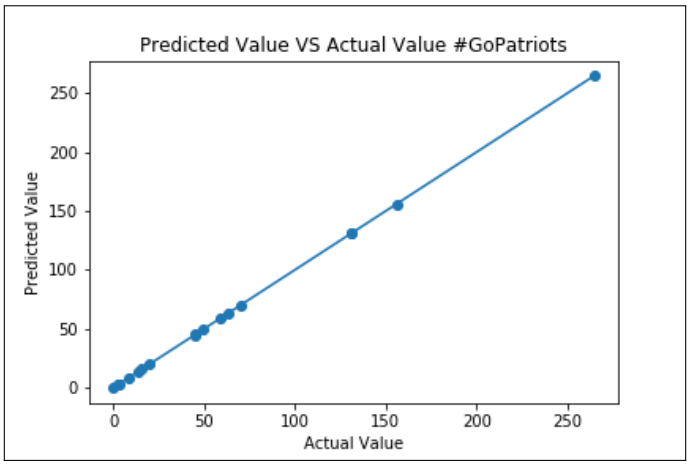


Figure 41: Predicted Sum of Sentiment Against Actual Sum of Sentiment

#GoHawks

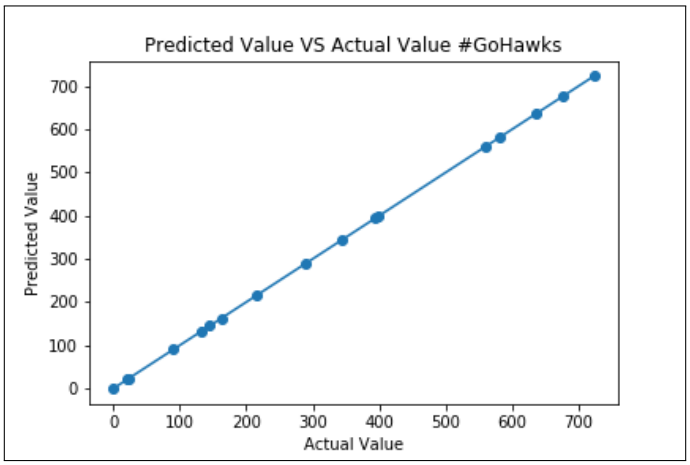


Figure 42: Predicted Sum of Sentiment Against Actual Sum of Sentiment

#NFL

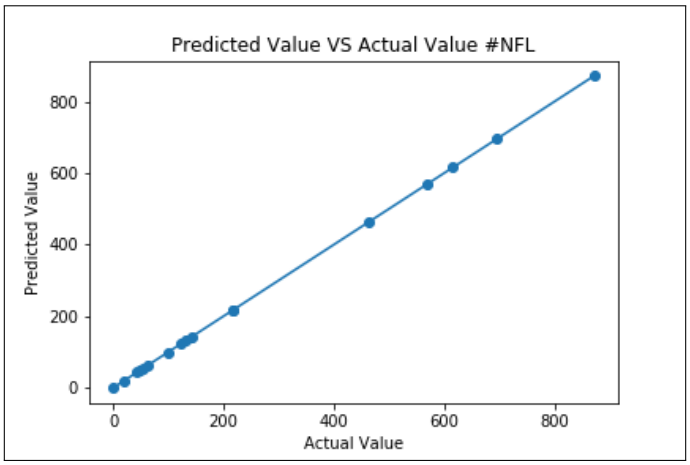


Figure 43: Predicted Sum of Sentiment Against Actual Sum of Sentiment

#Patriots

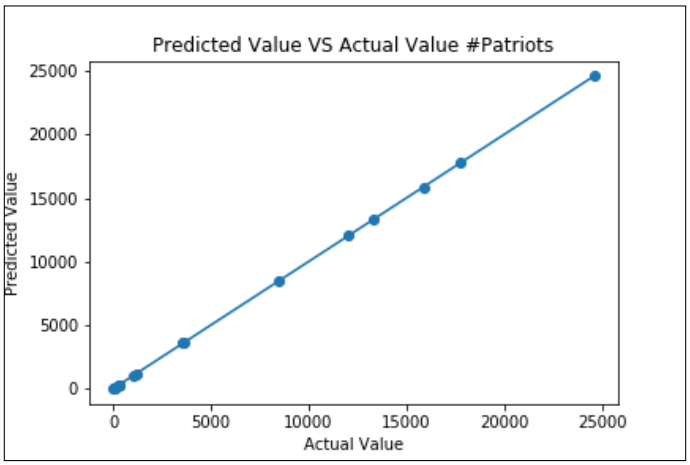


Figure 44: Predicted Sum of Sentiment Against Actual Sum of Sentiment

#SB49

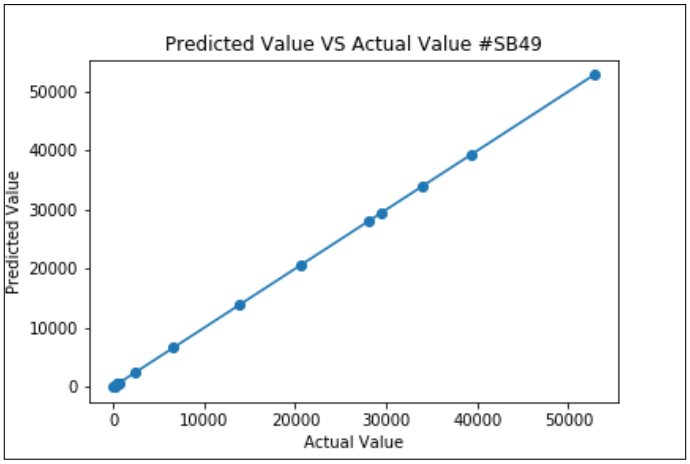


Figure 45: Predicted Sum of Sentiment Against Actual Sum of Sentiment

#SuperBowl

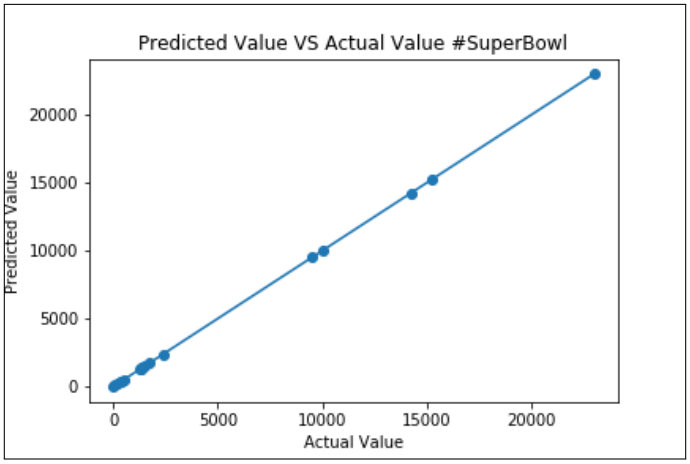


Figure 46: Predicted Sum of Sentiment Against Actual Sum of Sentiment

Hashtag	**Average RMSE**	R**-Squared Value**
#GoPatriots	1.25e-09	1
#GoHawks	5.33e-09	1
#NLF	1.17e-09	1
#Patriots	2.90e-07	1
#SB49	8.30e-06	1
#SuperBowl	4.85e-07	1

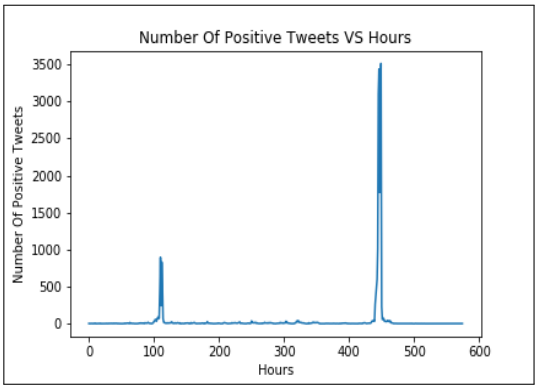
Table 21: Sentiment Predictor Performance for Different Hashtags

For every hashtag, the predicted value for the sum of the sentiment value matches the actual value of the sum of the sentiment value. This is a remarkable result because what it means is that the average sentiment of the tweets for the next hour can be predicted perfectly based on the average sentiment of the previous

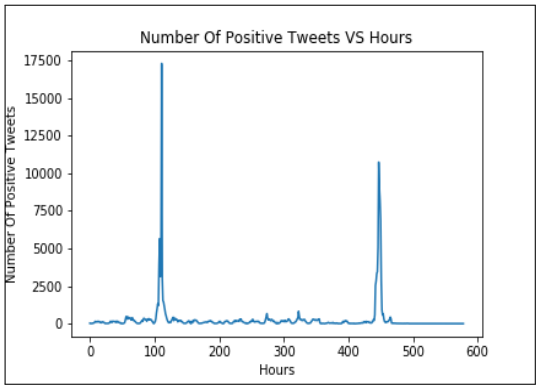
hour's tweets. It could be used to prevent people from using a particular hashtag if it is predicted that the next hour's tweets will have a very negative sentiment to them. For example, parents can safeguard their children from reading certain tweets, lessening the negative influence of social media on kids.

4.4.2 Team Performance from Tweet Sentiment

In this section, the performance of each team is predicted based on the number of positive and negative tweets, as well as the average sentiment.

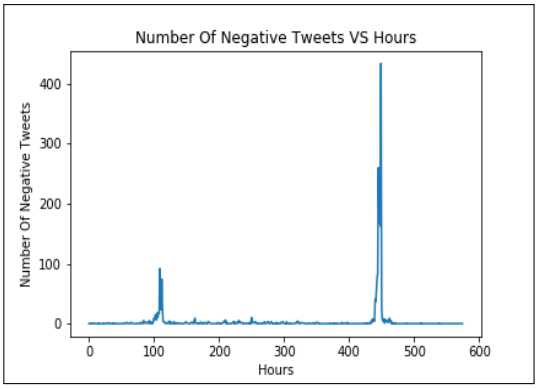


(a) #GoPatriots

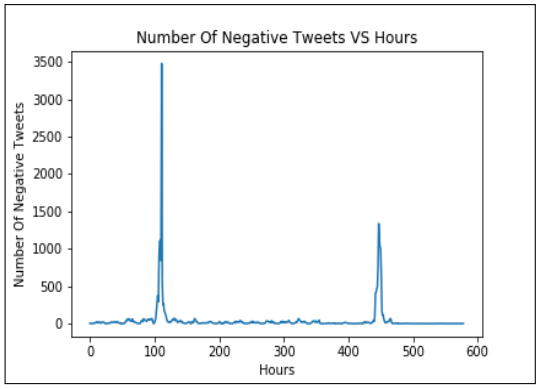


(b) #GoHawks

Figure 47: Hourly Positive Tweets for Entire Period

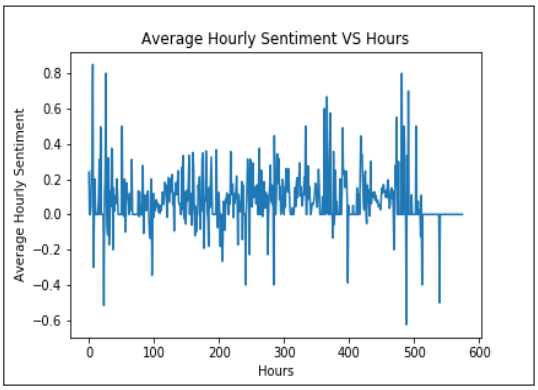


(a) #GoPatriots

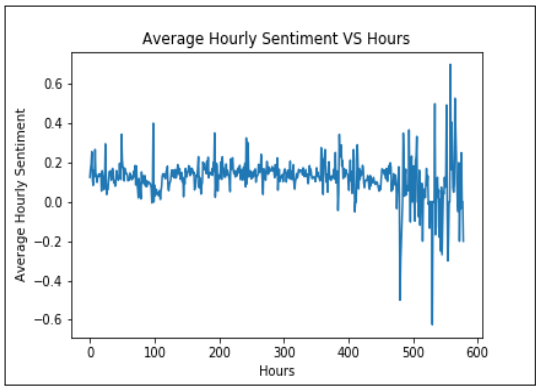


(b) #GoHawks

Figure 48: Hourly Negative Tweets for Entire Period



(a) #GoPatriots

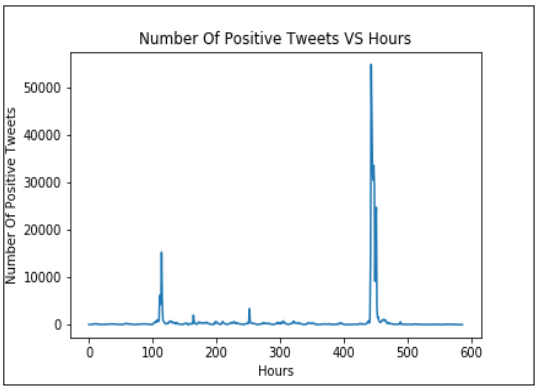


(b) #GoHawks

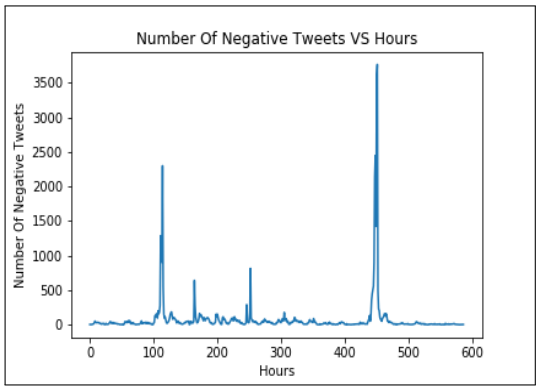
Figure 49: Average Sentiment for Entire Period

Looking at the average sentiment graphs for both hashtags, the first peak observed probably indicates that the teams qualified for the Super Bowl during this time period. The number of positive tweets for #GoHawks is almost 18 times that of #GoPatriots but the negative tweets are 35 times as much. This could probably mean that the Hawks qualifying for the Super Bowl was less expected and they beat the opponent team by a close margin. The second peak indicates when the Super Bowl was held. Both teams received a lot of positive support, with almost 10 times the negative sentiments from users.

To investigate further into the the rivalry and to see who won, the #Patriots hashtag is analyzed.



(a) Positive Tweets



(b) Negative Tweets

Figure 50: Hourly Tweets for #Patriots

It is clear that the #Patriots hashtag is more commonly used than #GoPatriots. During the Super Bowl, the positive support in the number of tweets for #Patriots is near the 50,000 tweet mark, hence it is hypothesized that the Patriots won. The final results confirm that the hypothesis is true.

4.4.3 Advertisement Sentiment

In this section, the sentiment for different advertisements aired during the Super Bowl is investigated and compared. The sentiments were collected through the tweets for the following companies:

1. Budweiser
2. BWM
3. Kia
4. Toyota
5. Snickers
6. T-Mobile
7. Mountain Dew
8. McDonald's
9. Budlight
10. Skittles
11. Nissan
12. Doritos

Choosing popular commercials ensures that the average sentiment is more representative due to the larger sample set. The data used is within the 17-hour period previously mentioned. The average sentiment, number of positive tweets and number of negative tweets for each commercial is tabulated below.

Company	**Avg Sentiment**	**Number of Positive Tweets**	**Number of Negative Tweets**
Budweiser	0.12	477	82
BMW	0.24	584	42
Kia	0.11	153	17
T-Mobile	0.011	502	42
Mountain Dew	0.17	35	8
McDonald's	0.199	191	25
Skittles	0.077	1660	272
Nissan	0.095	333	95
Doritos	0.17	405	49

Table 22: Sentiment for Different Advertisements

From the table above, it can be seen that the company with the highest average sentiment is BWM and the company with the lowest average sentiment is T-Mobile, which had an almost neural sentiment value even though the number of positive sentiments was 502. This suggests that there were some strong negative tweets about the T-Mobile commercial.

This information can, for example, be used by companies to find out who are their biggest supporters, or look at the negative tweets to find out what they can improve on. This is a very powerful tool as companies often pay millions of dollars to obtain such information by conducting surveys, which may not even be reliable.